

ABSTRACT

Title of Dissertation: DESIGN AND EFFECTIVENESS OF
MULTIMODAL DEFINITIONS IN ONLINE
SURVEYS

Maura Taylor Spiegelman
Doctor of Philosophy, 2020

Dissertation directed by: Professor Frederick Conrad
Joint Program in Survey Methodology
Michigan Program in Survey Methodology

If survey respondents do not interpret a question as it was intended, they may, in effect, answer the wrong question, increasing the chances of inaccurate data.

Researchers can bring respondents' interpretations into alignment with what is intended by defining the terms that respondents might misunderstand. This dissertation explores strategies to increase response alignment with definitions in online surveys. In particular, I compare the impact of unimodal (either spoken or textual) to multimodal (both spoken and textual) definitions on question interpretation and, indirectly, response quality. These definitions can be further categorized as *conventional* or *optimized* for the mode in which they are presented (for textual definitions, fewer words than in conventional definitions with key information made visually salient and easier for respondents to grasp; for spoken definitions, a shorter, more colloquial style of speaking). The effectiveness of conventional and optimized

definitions are compared, as well as the effectiveness of unimodal and multimodal definitions.

Amazon MTurk workers were randomly assigned to one of six definition conditions in a 2x3 design: conventional or optimized definitions, presented in a spoken, textual, or multimodal (both spoken and textual) format. While responses for unimodal optimized and conventional definitions were similar, multimodal definitions, and particularly multimodal optimized definitions, resulted in responses with greater alignment with definitions. Although complementary information presented in different modes can increase comprehension and lead to increased data quality, redundant or otherwise untailored multimodal information may not have the same positive effects. Even as not all respondents complied with instructions to read and/or listen to definitions, the compliance rates and effectiveness of multimodal presentation were sufficiently high to show improvements in data quality, and the effectiveness of multimodal definitions increased when only compliant observations were considered.

Multimodal communication in a typically visual medium (such as web surveys) may increase the amount of time needed to complete a questionnaire, but respondents did not consider their use to be burdensome or otherwise unsatisfactory. While further techniques could be used to help increase respondent compliance with instructions, this study suggests that multimodal definitions, when thoughtfully designed, can improve data quality without negatively impacting respondents.

DESIGN AND EFFECTIVENESS OF MULTIMODAL DEFINITIONS IN
ONLINE SURVEYS

by

Maura Taylor Spiegelman

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

Advisory Committee:

Professor Frederick Conrad, Chair
Professor Steven Heeringa
Professor Rochelle Newman
Professor Stanley Presser
Professor Michael Schober

© Copyright by
Maura Taylor Spiegelman
2020

Acknowledgements

Thank you to my committee members for your support, patience, and flexibility throughout this process. In particular, my chair, Frederick Conrad, helped this project grow from an offhanded musing into a research design, promising data, and exciting findings. Your insight, encouragement, and compassion were invaluable. I'm fond of this work, too.

Data collection was funded by the Rensis Likert Fund in Research in Survey Methodology, without which this project would not be possible. I'm grateful for the gentle pestering from my colleagues at NCES, even if my interest was piqued by a research topic that didn't fully overlap with my day job.

Thank you to my parents, friends, and family, and other cheerleaders along the way. And, to Edie, for fulfilling her role as personal assistant with helpful tasks such as creative typing and sitting on loose papers.

Table of Contents

Acknowledgements	ii
Table of Contents	iii
List of Tables	v
List of Figures	vi
Chapter 1: Introduction and background	1
1.1 Survey definitions	2
1.1.1 Definition comprehension	5
1.2 Past research on definitions in survey data collection	7
1.2.1 Research on definition accessibility	7
1.2.2 Research on definition placement	10
1.3 Multimodal communication	11
1.3.1 Working memory	12
1.3.2 Redundancy effect	14
1.3.3 Multimodal survey research	17
Chapter 2: Experimental Design and Data Collection	20
2.1 Study design	20
2.2 Hypotheses and aims	23
2.3 Data collection	26
2.4 Data cleaning	33
2.5 Respondent demographics	37
Chapter 3: Compliance with definitions	42
3.1 Compliance with spoken definitions	42
3.2 Compliance with textual definitions	47
3.3 Compliance with multimodal definitions	50
3.4 Self-reported compliance	53
3.5 Order of multimodal components	56
3.6 Reasons for compliance	57
3.6.1 Textual definitions	58
3.6.2 Spoken definitions	59
3.6.3 Multimodal definitions	62
3.7 Predictors of compliance	63
3.7.1 Gender	64
3.7.2 Age	65
3.7.3 Race and ethnicity	65
3.7.4 Education	66
3.7.5 Technology use	67
3.8 Summary	68
Chapter 4: Impact of definitions on data quality	71
4.1 Definition mode	71
4.1.1 Overall alignment with definitions	72
4.1.2 Alignment as a function of compliance	75
4.1.3 Intent to treat	79
4.2 Optimization	82

4.2.1 Overall alignment with definitions	83
4.2.2 Alignment as a function of compliance	85
4.2.3 Intent to treat	89
4.3 Definition inclusivity	90
4.4 Respondent characteristics.....	92
4.4.1 Gender.....	92
4.4.2 Age.....	93
4.4.3 Race and ethnicity.....	93
4.4.4 Education	94
4.4.5 Technology use	94
4.5 Summary	95
Chapter 5: Respondent burden and satisfaction	97
5.1 Respondent burden.....	97
5.1.1 Time to complete	97
5.1.2 Self-reported burden	100
5.2 Satisfaction.....	102
5.3 Summary	103
Chapter 6: Conclusion	105
6.1 Summary	105
6.2 Future research.....	107
Appendix A- Questionnaire	110
Appendix B- Mean Z-score by question and definition mode and optimization	130
Appendix C- Mean response by question and definition mode and optimization	132
Bibliography	134

List of Tables

Table 1- All possible definition treatment groups	21
Table 2- Experimental definition treatment groups	23
Table 3- Data collection schedule	29
Table 4- Flesch-Kincaid reading level and estimated reading and listening time, by definition optimization, mode and question.....	32
Table 5- Sample sizes by definition treatment.....	36
Table 6- Distribution of responses by technology-based activity (How often do you...)	40
Table 7- Compliance with spoken definitions by question number and definition type	43
Table 8- Percentage distribution of number of questions for which a respondent fully played definitions by definition type	46
Table 9- Compliance with textual questions and definitions by question number and definition type	48
Table 10- Percentage distribution of number of questions for which text could be fully read by respondent and definition type	49
Table 11- Compliance with multimodal definitions by question number and optimization	51
Table 12- Percentage distribution of number of questions for which definitions were listened to and questions and definitions could be fully read by respondent and definition type	52
Table 13- Self-reported compliance with definitions by respondent and definition type	54
Table 14- Self-reported ordering of multimodal definitions.....	56
Table 15- Mean Z-score by definition mode	73
Table 16- Mean Z-score by definition mode for compliant observations	76
Table 17- Mean Z-score by definition mode received.....	80
Table 18- Mean Z-score by definition mode and optimization	84
Table 19- Mean Z-score by definition mode and optimization for compliant observations	85
Table 20- Mean Z-score by definition mode received and optimization.....	89
Table 21- Time spent on 12 definition questions (in seconds)	98
Table 22- Comparison of questionnaire timing by treatment	99
Table 23- Self-reported burden by definition mode and optimization	101
Table 24- Respondent satisfaction by definition mode and optimization.....	102

List of Figures

Figure 1- Mean Z-score by definition mode and compliance status.....	78
Figure 2- Mean Z-score by definition mode and optimization and compliance status	88

Chapter 1: Introduction and background

Ensuring that survey respondents and researchers hold a shared understanding of terms used in survey questions is crucial for maintaining data quality. When respondent and researcher intentions are misaligned, respondents may provide different information than researchers intended, resulting in inaccurate reporting. This agreement can be achieved with the use of definitions, as explicitly clarifying terms can provide conversational grounding between respondents and interviewers or, for self-administered surveys, between respondents and researchers (Clark 1996). Past research on the delivery of definitions in survey data collection has explored their availability and accessibility (whether presented to all respondents or subgroups based on explicit requests or behavioral indications of comprehension difficulty or misunderstanding (Conrad et al. 2006; Peytchev et al. 2010; Zhang and Conrad 2014)), standardization (for interviewer administered surveys, whether a script is followed strictly, i.e., interviewers cannot provide clarification if it is not provided to all respondents. or whether interviewers have the autonomy to provide additional clarification to respondents (Schober 2000)), and placement (for visual surveys, where to place definitions such that they will be seen and used by respondents (Redline 2013; Metzler, Kunz, and Fuchs 2015)). This prior research on surveys concerns the delivery of unimodal, that is, solely spoken or solely textual definitions. Other fields, particularly educational psychology, have focused on multimodal communication and found that this can improve comprehension and information retention over unimodal communication. This dissertation builds upon both of these

research domains to explore whether multimodal definitions, that is, presenting both spoken and textual clarification, can improve data quality relative to unimodal or no clarification. The conditions under which multimodal definitions might be effective are evaluated, in particular optimization (whether definitions are identical in wording for spoken and textual presentations or designed to exploit the particular affordances of the different modes to best communicate the designers' intentions). Finally, this research explores respondents' compliance and satisfaction with these new formats.

A discussion of past research is divided into two primary sections: (1) survey definitions and (2) multimodal communication, i.e., not about definitions per se. The section on definitions emphasizes the importance of ensuring respondents understand researchers' intentions, the circumstances under which respondents request or read definitions, and placement of definitions. The section on multimodal communication discusses the processing of textual and spoken communication, the impact of working memory on this communication, effectiveness of presenting multimodal instructional information on educational outcomes, and applications to survey methods. These provide justification for the research design and hypotheses described in Chapter 2: Experimental Design and Data Collection.

1.1 Survey definitions

Surveys ask respondents about conditions or situations of varying complexity, clarity, and familiarity to respondents. For ideas or terms for which respondents' understanding is different from researchers' intentions, standardizing understanding is key for maintaining data quality (Conrad and Schober 2000; Schober, Suessbrick, and

Conrad 2018). Differences between researchers' intentions and respondents' understanding can introduce measurement error when respondents' interpretations of survey terms do not align with researchers' intended meanings or when misalignments occur among key population groups. For example, the concept of who lives in a household is straightforward in most living situations. Misaligned understandings will not matter for this majority of households, since a definition is likely to contain information that is irrelevant to prototypical households and thus how they respond to the question. However, for people with complicated situations, such as a child living away at college on the reference date (or, perhaps more timely, a child who typically lives away from college but is at her parents' household after her in-person classes transitioned to online classes) – should she be counted as living in the household or not? – misalignments can result in increased variance or, if misalignments occur in one direction, bias (Tourangeau et al. 2006; Conrad and Schober 2000). Many respondents have no need to consider whether to count children away at college as members of their household, but instructions matter for the portion of respondents who have college-aged children who live at or near their school. The impact of these instructions depends on the prevalence of the misaligned characteristic (here, college-aged children not living at home) and how closely researchers' definitions align with respondents' understanding of terms (whether respondents would include that child as a household member if no definition of *household member* were available). Misaligned understandings can result in a systematic over- or under-count of household size, depending on the definition of

who should and should not be included. Under-counted and over-counted individuals may be different on key measures than people with straightforward living situations.

While definitions can improve response quality when they help align respondents' interpretations with researchers' intentions, there are potential drawbacks of presenting this additional information. Providing definitions may increase the amount of time needed to complete survey items (Schober and Conrad 1997; Schober, Conrad, and Fricker 2004; Conrad and Schober 2000). When respondents were shown unnecessary definitions, that is, definitions that matched respondents' understanding of everyday terms, Yan (2005) found only weak evidence that response time increased, though cautioned that increased response variance may be another negative result of presenting irrelevant definitions (that did not apply to respondents) or technical definitions (regardless of whether they provided new information beyond respondents' understanding of common terms). Conrad, Schober, and Coiner (2007) similarly found that respondents who viewed definitions in a web survey, either by deliberately clicking a link or having definitions automatically displayed when they were slow to respond, took longer to complete survey items than respondents who did not view definitions. Study participants reported that they liked the ability to receive clarifying definitions, so the increased time to request and read definitions, which could increase respondent frustration or perceived survey burden, was not necessarily seen as a drawback by respondents. It is possible that for other definition formats (for example, multimodal definitions), respondents may perceive them to be burdensome, so burden, both inferred and self-reported, must be considered when evaluating best

practices. Irrelevant definitions may not affect data quality (Conrad et al. 2006), but respondent perceptions should be considered when using a new type of definition, such as multimodal presentation in a typically visual mode.

When definitions are displayed by default in textual questionnaires, rather than being available upon request or targeted based on response time or other behaviors, these definitions can potentially overwhelm respondents and make it difficult to discern relevant information. Consequently, respondents may simply opt not to read definitions when they are always provided. However, this concern has little empirical support. For example, Peytchev et al. (2010) found that responses from respondents who were always shown definitions in a web survey differed from responses from those who were not able to access definitions (with responses from the definition group presumably more aligned with definitions). Therefore, it is likely that “always-on” definitions present some benefits to data quality, and there is no reason to believe that omitting definitions entirely is preferable to showing definitions by default to all respondents.

1.1.1 Definition comprehension

Respondent use of definitions involves several necessary steps: respondents must be exposed to definitions, acquire the information presented, and integrate the definition into their responses. If definitions are presented by default, respondents must listen to or read that information, understand it, and incorporate it into their responses. If definitions are only presented upon request, exposure is contingent upon respondents first deciding they must seek out a definition by asking an interviewer or taking

deliberate steps (for example, clicking a link with their mouse or looking at the cover page of a paper questionnaire).

It is important to note that while much of the motivation for this dissertation is based on work in educational psychology that directly assesses comprehension through measures such as the accuracy of answers to test questions (see section 1.3 Multimodal communication), the current study assumes respondents comprehended definitions if they provide responses that align with definitions. Comprehension was not directly measured with a post-survey quiz or exam (which has been done in a number of published lab-based studies, e.g., Suessbrick, Schober, and Conrad 2000; Schober, Suessbrick, and Conrad 2018), because respondents are typically expected to retain comprehension of survey definitions for a few survey items, rather than an entire questionnaire.. Instead, for this dissertation, definitions were designed as either *inclusive* (prompting higher numerical counts from respondents when used) or *exclusive* (prompting lower numerical counts from respondents when used). As further discussed in 2.3 Data collection, where responses differ by the type of definition shown to respondents, it can be inferred that higher numerical responses to questions with inclusive definitions and lower numerical responses to questions with exclusive definitions can be attributed to definition use. While comprehension is not measured directly, definition use assumes that respondents first comprehended definitions, and then integrated that information into their numerical responses.

1.2 Past research on definitions in survey data collection

For primarily visual surveys, that is, surveys administered by paper or online, definition accessibility (whether presented to all respondents or subgroups based on explicit requests or behavioral indications of comprehension difficulty or misunderstanding) and placement (where to place definitions such that they will be seen and used by respondents) are two lines of research. This dissertation study does not vary definitions' accessibility; all textual definitions are persistent and shown by default, while all spoken definitions require deliberate action from respondents (clicking "play"). Similarly, the location of definitions does not vary in the current study, but is based on prior research on the frequency and conditions in which respondents access and use definitions, as well as typical web survey and online practices.

1.2.1 Research on definition accessibility

Before a respondent can comprehend a definition, they must first be presented with that definition. This often involves some deliberate action from a respondent, though the level of effort involved may vary. Examples of the lowest effort from respondents are situations in which definitions are presented to everyone, for example, in-person interviewers reading definitions aloud to all respondents or a web survey in which textual definitions are shown to all respondents. More often, a higher level of effort is required. For example, in a paper survey, respondents may have to turn to a separate page of the questionnaire. In a telephone interview, a respondent may have to ask an interviewer to provide a definition. In a web survey, a respondent may have to click a link or roll their cursor over key terms. In these situations, respondents are unlikely to

invest great effort to access definitions, and even low-effort actions are avoided by respondents. For example, in web surveys in which clarifying text is not automatically shown to all respondents, respondents are unlikely to view a definition through moderate-effort actions such as clicking a mouse or even low-effort actions such as rolling a mouse over the term to be defined, though they are more likely to take low-effort actions than to complete relatively more burdensome steps, such as making a single or multiple clicks (Conrad et al. 2006). Respondents' reluctance to obtain online definitions persists even when given specific instructions or information about the usefulness of low- and moderate-effort definitions (Peytchev et al. 2010; Conrad, Schober, and Coiner 2007).

An eye-tracking study revealed that not only are respondents unlikely to click or roll their mouse over key terms to be presented with definitions, respondents are unlikely to read the definitions they are shown (Galesic et al. 2008). This study found that about 10 percent of study participants both rolled their mouse over a key term and looked at the definition, while about 78 percent of participants to whom definitions were shown with no required effort looked at each definition. However, respondents who requested definitions through mouse rollovers spent more time looking at them than respondents for whom they appeared automatically, and increased time looking at definitions (regardless of whether they were automatically shown or respondent-requested) was linked to increased response accuracy. The relatively few respondents who request definitions seem to provide high quality data, but this effect may be

dwarfed by the infrequency with which the majority of respondents request definitions or do not appear to read persistent textual definitions.

Respondent reluctance to expend effort in reading or even accessing definitions is difficult to overcome. For example, demonstrating the potential usefulness of definitions before a survey begins has been unable to increase respondents' definition access. Some respondents may be unaware that a definition is relevant to their situation or believe they understand key survey terms, not realizing that requesting a definition could improve the accuracy of their responses. However, even if respondents are reminded that their understanding of terms may be different from researchers' understanding, they are still reluctant to even roll their mouse over key terms (Peytchev et al. 2010). In that study, respondents began a web survey by being trained on definitions, specifically, reading a definition, reporting what was surprising about that definition, and indicating their familiarity with technical terms used in the definition, in order to convey the message that requesting and reading definitions was a worthwhile activity. Respondents who received this training were actually less likely to request definitions than those who were not trained, and there is minimal evidence to suggest that respondents trained to request definitions were more likely to incorporate definitions into their responses.

In this research, definition accessibility is not further tested experimentally. Given the consensus findings of past studies, textual definitions were presented to all respondents who are assigned to experimental conditions with textual definitions,

rather than available only upon demand, in order to maximize the number of respondents exposed to these definitions. However, because this past research focused only on textual definitions, it could not be assumed that respondents would react similarly to persistent spoken definitions. For this dissertation, minimal effort was needed to play spoken definitions (a single mouse click), under the assumption that respondents would react poorly to sound that automatically began to play without a user-initiated action.

1.2.2 Research on definition placement

Respondent exposure to definitions also depends on definition placement. When definitions are automatically presented to all respondents, rather than available upon request or via deliberate mouse actions, definition location should be a key decision in online survey design. When presented visually, definitions should be able to be located with minimal effort, noticeable to respondents, and near relevant questions. In web surveys, respondents have been found to use definitions more frequently when they appear immediately before questions, rather than immediately after (Redline 2013). Respondents fixate on definitions more frequently and for longer durations when placed before a question than when they appear either between a question and response options or with response options (Kunz and Fuchs 2012). However, when instructions are visually distinct from question stems, instructional definitions are used more frequently when they appear between a question and response options, rather than before a question (Metzler, Kunz, and Fuchs 2015). Follow-up studies have not yet revealed whether placement (before or after), distance (space between

question and definition or no space), or an interaction between these factors (Redline et al. 2016) drives respondent behavior.

Definition placement does not vary in this research. Instead, textual definitions appear in the same location for all respondents: visually distinct from question stems, and between question stems and response boxes, in order to maximize their visibility when shown to respondents (and media bars to play spoken definitions appear in the same location). It should be noted that while survey research on definition placement has focused on visual modes, the under-researched equivalent question for spoken definitions is whether they are used more frequently when presented before or after the question to which they apply. The current study uses spoken definitions and written questions, rather than spoken questions, so the ideal relative placement of spoken questions and spoken definitions (for example, in a telephone survey) remains an open question.

1.3 Multimodal communication

Multimodal communication involves multiple strategies through which information is conveyed. In surveys, this can apply to both output, that is, the information presented to respondents, and input, that is, the information reported by respondents (Johnston 2008). For example, respondents can be exposed to multimodal output by both viewing a written question and hearing it read aloud. Respondents can have the option of multimodal input by being permitted to speak their response or enter a corresponding number on a telephone keypad (similar to touchtone data entry or TDE). This study focuses on multimodal output, that is, conveying information to

respondents through multiple channels. The following sections discuss memory and cognition, multimodal learning, and multimodal survey methods in order to support the hypothesis that multimodal and mode-optimized strategies can improve the comprehension and use of survey definitions.

1.3.1 Working memory

Working memory is a key component of the survey response process. Respondents use their working memory in order to understand questions being asked, by maintaining words read or heard earlier while still perceiving later words (Tourangeau, Rips, and Rasinski 2000). Baddeley's model of working memory specifies three components: central executive, phonological loop, and visuospatial sketchpad (Baddeley 1992). The central executive system controls attention, while the phonological loop concerns verbal information, regardless of communication channel; verbal information conveyed either aurally or visually is maintained in the phonological loop. Finally, the visuospatial sketchpad processes visual information including color, movement, and shape. While these processes are related, one of the key pieces of Baddeley's model is the partial independence of the phonological loop and visuospatial sketchpad. That is, because these processes do not wholly occupy the same processors, an individual can use more working memory when both the verbal and visual components are engaged than when information is conveyed using only one of these components (Baddeley 1992; Dumas, Lalanne, and Oviatt 2009). Even

when information is presented simultaneously, more can be processed when different components are used.

Educational psychology research supports this theory, sometimes referred to as the modality effect. For example, Moreno and Mayer (2002) provided participants with verbal information about lightning either aurally-only or simultaneously both aurally and visually. Following this information, short exams were administered to measure participant comprehension. When comprehension was measured on three different dimensions (information retention, for which participants were asked to write, in their own words, what they learned; matching, for which participants were asked to interpret diagrams related to the lightning process; and transfer, for which participants were asked to apply the information they learned in a new context by demonstrating cognitive processing and creativity, rather than simply recalling literal content), multimodal participants demonstrated higher comprehension than aural-only participants on all three dimensions. In addition, Mousavi, Low, and Sweller (1995) operationalized comprehension as the amount of time needed for participants to solve geometry problems, with less time indicating higher levels of comprehension. Study participants were shown diagrams as well as textual-only, aural-only, or both textual and aural verbal instructions. However, they found that their subjects performed best, that is, needed less time to accurately solve geometry problems related to the presented concepts, when diagrams were combined with aural verbal information only (compared to textual only or both aural and textual). By comparing sequential and simultaneous presentation, they attributed these results to the relatively low

cognitive load of using multiple communication channels, due to partial independence of visual and verbal processing. That is, presenting simultaneous, identical verbal and visual information can improve comprehension unless paired with additional stimuli that overburden participants.

In survey research, the working memory model is particularly applicable to interactions that are only aural. While respondents have the option of asking interviewers to repeat information, they are primarily relying on their short-term recollection. When surveys are partially or completely visual, respondents can easily refer back to persistent questions or response options, resulting in a process that is less cognitively taxing and less reliant on working memory. For unimodal definitions, this theory suggests that unimodal textual definitions are less burdensome to respondents than unimodal spoken definitions. For multimodal definitions, this theory suggests that presentation of the same information in different modes, in particular, both presenting text to respondents and having respondents hear the same text read aloud, is less cognitively burdensome to respondents than a single mode. Further, multimodal presentation has the potential to improve data quality for complicated concepts and definitions if respondents learn best from multimodal stimuli.

1.3.2 Redundancy effect

Research on multimodal communication has combined spoken information with a variety of visual presentations, building upon the working memory model and cognitive load theory. Studies have generally found that combining both audio and visual material is more effective than using only one mode; for example, in

educational psychology, researchers have found that presenting students with both audio and visual material is a more effective teaching method than using only spoken communication (Mousavi, Low, and Sweller 1995; Moreno and Mayer 2002). However, this effect is most often explored with non-text based visual stimuli, such as diagrams or drawings. When verbal information is simultaneously conveyed both aurally and textually, redundancy can potentially lead to reduced comprehension when stimuli are designed to be processed simultaneously. For example, when an animated explanation of lightning was combined with either spoken narration only or identical, simultaneous spoken and textual narration, the treatment of identical, simultaneous text and audio resulted in poor comprehension, as operationalized by information retention and transfer (Mayer, Heiser, and Lonn 2001). That is, participants who were exposed to redundant spoken and written text and a complementary animation had lower comprehension than participants who were exposed to only spoken text and a complementary animation. Note that many of these studies involved visual stimuli that were difficult to cognitively process, such as combinations of written instructional text, numerical tables, and graphs or diagrams from Kalyuga, Chandler and Sweller (2004). Designed to exceed study participants' maximum cognitive capacity, redundant stimuli did not improve outcomes, though the visual animation may have also played a role in the results. However, most survey

researchers hope to convey slightly less complicated material to respondents, without the same goal of exceeding cognitive capacity.

For information of any complexity level, stimuli that present the same underlying message without solely reading text aloud are not necessarily subject to a redundancy effect (Kalyuga, Chandler, and Sweller 2004). Mild levels of redundancy, for example, key words or phrases only, have been shown to increase retention (Mayer and Johnson 2008). Similarly, presenting survey respondents with key information from definitions only, rather than printing an interviewers' (relatively lengthy) script, are expected to increase comprehension or, at a minimum, not reduce comprehension. That is, redundancy can yield higher rates of comprehension when key pieces of narration are emphasized, rather than simply duplicating spoken information. More specifically, presenting identical spoken and textual definitions could reduce comprehension, while complementary definitions, designed specifically for their spoken or textual delivery, may potentially improve comprehension. This study tests two types of multimodal definitions: one with fully redundant spoken and textual information (i.e., the same text shown visually and read aloud) and one with redundant key points only (i.e., the same concepts conveyed textually and aurally). If independent channels drive the effectiveness of multimodal communication, then both types of multimodal definitions are predicted to outperform unimodal definitions. If the redundancy effect is at play, then complementary multimodal

definitions are predicted to outperform both redundant multimodal and unimodal definitions.

1.3.3 Multimodal survey research

Most survey modes employ a single communication channel through which questions are posed to respondents. For example, survey questions are only presented aurally in telephone interviewing and questions are only presented textually in paper surveys.

While the channel through which respondents provide their answers is often the same channel in which questions are presented, this is not a requirement (Couper 2011).

For example, a web survey may pose questions to respondents aurally via pre-recorded videos but require respondents to select responses from a written list.

There are exceptions to this pattern of single-channel modes, and some survey designs employ multi-channel interviews in which information is presented to respondents both aurally and textually. For example, ACASI (audio computer-assisted self interviewing) typically shows questions and response options on a computer screen while a voice reads the same material aloud. This combination is most often used for questions on sensitive topics, providing privacy to respondents and yielding higher reported rates of socially undesirable behavior than interviewer-administered interviews (Tourangeau and Smith 1996). Isolating the impact of adding spoken communication to a visual survey (that is, comparing CASI to ACASI) is challenging due to respondent compliance; in the field, respondents often ignore the aural component by reading questions and responses to themselves more quickly than they can be read aloud, reducing the experience to a primarily visual survey mode. In

addition, many respondents simply decline to wear the provided ACASI headphones, producing an experience without any aural component (Couper, Tourangeau, and Marvin 2009).

Survey data collected by CAPI (computer-assisted personal interviewing) may use showcards, such that an interviewer both reads response options aloud and shows respondents a list of printed answers. While ACASI is generally used to reduce an interviewer's social presence, showcards typically display response options, rather than question text or supplementary information, with the goal of reducing the cognitive burden of a completely aural process (Rogers 1976). More recently, Jäckle, Roberts, and Lynn (2006) measured the effect of CAPI showcards on respondent satisficing. When response options were presented both aurally and on showcards or aurally only, there were minimal differences in non-differentiation and response order effects. However, ordered and repeated response options, such as a Likert scale that applies to a series of questions, may present less of a cognitive burden than potentially lengthy definitions that contain surprising or unexpected details, and the quality of respondents' answers could well benefit from the use of multimodal definitions.

A few tactics have been used to promote both aural and visual channels in telephone interviews, even without an in-person interviewer to hold the appropriate showcard or other visual aid. For example, printed materials have been mailed to respondents to retrieve and consult during an interview, and some respondents have been asked to

draw their own showcards, though these strategies are not routinely employed (Miller 1984). Advances in technology and increased respondent use of technology present a new array of strategies to convey information to respondents, moving beyond mailed supplements and hand-drawings. Consequently, best practices should be evaluated and updated to reflect these new technological developments and uses (Schober and Conrad 2008). For example, telephone interviews with respondents on smartphones can involve traditional spoken interviews or be combined with SMS texting, web access, or video communication. Over 80 percent of American adults own smartphones, with higher penetration among certain subgroups such as young adults (Anderson 2019), and techniques that take advantage of device capabilities warrant further research. This study isolates the effects of different multimodal communication characteristics in order to measure their effect and determine the combination that produces the highest quality data from survey respondents.

Chapter 2: Experimental Design and Data Collection

This chapter introduces and justifies the experimental design used for this research.

The chosen set of definition presentation formats are tied to research hypotheses.

Data collection procedures are explained, as well as the procedures for cleaning data and detecting suspicious or otherwise outlying observations. Finally, the demographic characteristics of respondents are presented and compared across each definition treatment.

2.1 Study design

This research is intended to explore whether multimodal definitions, that is, presenting both spoken and textual information, can improve data quality relative to unimodal clarification. For this study design, two additional factors were considered: the accessibility of definitions (whether presented to all respondents or only to respondents who explicitly request a definition) and optimization (whether definitions are in conventional spoken and textual formats or designed to exploit the particular affordances of the different media to best communicate the designers' intentions). Varying all of these factors would result in 24 different experimental conditions, shown below in Table 1.

Table 1- All possible definition treatment groups

		Textual					
		None		Conventional		Optimized	
Spoken	None	C		TC-A	TC-D	TO-A	TO-D
	Conventional	SC-A	SC-D	SC-A*TC-A	SC-A*TC-D	SC-A*TO-A	SC-A*TO-D
				SC-D*TC-A	SC-D*TC-D	SC-D*TO-A	SC-D*TO-D
	Optimized	SO-A	SO-D	SO-A*TC-A	SO-A*TC-D	SO-A*TO-A	SO-A*TO-D
				SO-D*TC-A	SO-D*TC-D	SO-D*TO-A	SO-D*TO-D

Multimodal treatments are indicated with a *, and definition conditions are written with three letters separated by a dash. The first position indicates mode, with T for textual definitions and S for spoken definitions. The following letter indicates optimization of that mode, with C for conventional definitions and O for optimized definitions. Accessibility follows the dash; A indicates definitions that are provided automatically, and D indicates definitions provided upon demand only. For example, the cell “TC-A” indicates textual, conventional definitions that appear automatically and no spoken definitions, while the cell “SO-D*TO-D” indicates a multimodal treatment with optimized spoken and textual definitions, both available on demand. Finally, C denotes a control group that receives no definitions.

However, this study prioritizes experimental conditions that can evaluate (H1) whether response alignment with definitions can be improved with multimodal definitions and (H2) whether response alignment with definitions can be improved with definitions optimized for the mode(s) in which they’re presented, and the full set of 25 treatment groups are not needed to answer these questions. In particular, in

order to test the dueling theories of why multimodal definitions may be effective (see 1.3.2 Redundancy effect), all 16 possible multimodal treatments (varying conventional and optimized components, as well as whether components are available automatically on demand) are not necessary. Multimodal definitions that mix conventional and optimized components (for example, SO-A*TC-A) do not contribute to evaluating that theory and can be considered a low priority, as they do not address H2. In addition, it is well established that respondents may be unlikely to click on supplemental material (see 1.2.1 Research on definition accessibility). Given that prior work, further validation is a low priority for this research, but it can be used to give these treatments their “best chance” of succeeding. As a result, textual on-demand definitions can be eliminated.

Past research on definition accessibility has studied textual definitions, rather than spoken definitions. It remains an open question as to whether similar results would occur when comparing persistent and on-demand spoken definitions. Use of spoken definitions in an online survey is a new presentation format, but rather than maximizing their potential exposure with automatic play to all respondents, only on-demand spoken definitions were used in the current study, requiring respondents to click a “play” icon on a media bar. On-demand implementation of spoken definitions allows us to observe the types of respondents that actively play spoken definitions and the situations in which they are most likely to do this, since their behavior can be tracked with survey paradata. In addition, it was our intuition that respondents might react negatively to autoplay audio files, particularly if they were accustomed to loud,

autoplay advertisements on other websites they visited. However, systematically varying the accessibility of spoken definitions remains an open question that can be addressed in a follow-up study.

By removing treatments with automatic spoken definitions and multimodal treatments that combine conventional and optimized components, 7 treatment groups remain, as shown in Table 2: a control group with no definitions; textual conventional definitions (always on); textual optimized definitions (always on); spoken conventional definitions (on-demand); spoken optimized definitions (on-demand); multimodal conventional definitions (textual always on and spoken on-demand); and multimodal optimized definitions (textual always on and spoken on-demand).

Table 2- Experimental definition treatment groups

		Textual					
		None		Conventional		Optimized	
Spoken	None	C		TC-A		TO-A	
	Conventional		SC-D				
				SC-D*TC-A			
	Optimized		SO-D				
						SO-D*TO-A	

2.2 Hypotheses and aims

As discussed in 2.1 Study design, the experimental design prioritizes assessing the effectiveness of multimodal definitions relative to unimodal definitions (H1). In particular, the primary hypothesis is that multimodal definitions will yield higher quality data, that is, responses more consistent with definitions, than unimodal definitions. This design also tests whether any multimodal definition (conventional

and optimized) outperforms unimodal definitions, or if only multimodal optimized, rather than multimodal conventional definitions, lead to responses that are more consistent with definitions.

In addition to comparing conventional and optimized multimodal definitions, optimization will also be evaluated for unimodal definitions. In particular, it is expected that unimodal optimized definitions will yield responses more consistent with definitions than unimodal conventional definitions (H2). Conventional definitions were designed to follow the format of data collection instruments from many federal statistical agencies. These conventional definitions are informative, but, when presented in textual format, appear as a dense paragraph, and it may be difficult for respondents to identify the subcomponents most relevant to their situations. These same definitions, when read aloud, do not flow like a conversation or other spoken communication. This style also mimics some production environments, for example, a primarily self-administered paper or online questionnaire in which some respondents may complete the instrument over the phone while a telephone interviewer reads the self-administered questionnaire to respondents. For multimodal conventional definitions, identical wording was used for both the spoken and textual components, while for multimodal optimized definitions, spoken optimized and textual optimized components were presented together.

Optimized definitions were designed to be more appealing for respondents to read or listen to and easier for respondents to identify relevant information by following best practices of written and spoken communication. For textual optimized definitions, factors that facilitate text comprehension were used. In particular, bolded text was

used to draw attention to key words and phrases, and organizational devices such as bullets were used to divide text into logical groupings (White 2010).

For spoken optimized definitions, these scripts were designed to follow best practices for spoken communication. For example, spoken optimized definitions remove extraneous information that is included in their conventional counterparts, in order to facilitate comprehension (Sweller et al. 1990). Shorter spoken definitions are also less taxing on the working memory of respondents, and require a relatively lower cognitive burden to comprehend than longer, conventional definitions (Leahy and Sweller 2011). During the creation of spoken optimized definitions, these instructions were also read aloud, recorded, and played back to judge their flow and ease of comprehension, then adjusted, if needed, as in iterative process. However, they did not undergo external cognitive or usability testing.

Comparing unimodal spoken and textual definitions, it is expected that responses to textual definitions will be more aligned with definitions than responses to spoken definitions. Given the persistent nature of written text, respondents can easily refer back to them with minimal reliance on their working memory (Singh, Marcus, and Ayres 2012; Leahy and Sweller 2011). Presumably, this reduces the effort needed to use and apply textual instructions than spoken instructions to the survey process.

In addition, we will explore the extent to which respondents comply with instructions to read or listen to definitions. It is expected that compliance will vary by definition

type, and compliance is expected to be higher for optimized definitions than their conventional counterparts, since optimized definitions were designed to be easier to comprehend and, by removing extraneous information, involve less time to read or listen to than conventional definitions.

Finally, we will explore respondent burden and satisfaction. It is expected that respondents with spoken definitions (unimodal spoken and multimodal) will consider the survey to be more burdensome than unimodal textual respondents, that they will spend more time completing the questionnaire, and this increased burden may translate to decreased satisfaction with the survey process. Comparisons between unimodal textual and multimodal respondents are key for determining whether the additional stimuli and tasks asked of respondents (regardless of whether they do, in fact, fully comply with multimodal definitions) might serve as “deal breakers” for implementation, even when accounting for potential improvements in data quality with multimodal definitions.

2.3 Data collection

Questionnaires were programmed into Qualtrics for each of the 7 treatment groups (control, spoken conventional, spoken optimized, textual conventional, textual optimized, multimodal conventional, multimodal optimized). Spoken definitions appeared as audio files, and respondents could listen to these definitions by clicking the “play” icon on a standard media bar. The number of times each audio file was played in its entirety for each respondent was captured, as well as the elapsed time a respondent spent on each page of the questionnaire. Respondents were recruited from

Amazon Mechanical Turk (MTurk), using TurkPrime for respondent management. A \$1 incentive was provided to respondents upon completion of the survey, consistent with or slightly higher than payment rates for MTurk activities of similar length and burden. This study was approved by the University of Maryland Institutional Review Board.

The full survey instrument is shown in Appendix A- Questionnaire. Respondents were asked to provide numerical responses to 15 questions. For 12 of these questions, definitions were either “inclusive” (five questions) or “exclusive” (seven questions). The former were designed to elicit higher responses by expanding the set of eligible behaviors; the latter were designed to elicit lower responses by restricting the set of eligible behaviors. For 3 questions, definitions were designed to be neutral, that is, purely descriptive but irrelevant, without either expanding or restricting eligible behaviors. In addition, 3 questions with inclusive definitions asked respondents to list items by providing text responses, rather than numerical responses, in order to test the impact of experimental definitions on a different question type (i.e., questions that ask respondents to provide free text responses, rather than numerical responses). In the analyses presented here, note that text response questions were used when cleaning data (see 2.4 Data cleaning), but are not included in analyses of data quality.

To promote the questionnaire’s coherence, questions on similar topic areas were grouped together (for example, hours spent watching television and listening to the radio). Thus, all respondents viewed questions in the same order. Finally, all

respondents were asked a series of debriefing questions on demographics and their experience during the study.

Data were collected over 3 rounds, detailed below. Multiple rounds were used to verify that data collection procedures and survey instruments functioned as intended in production mode, that respondents played spoken components of definitions, and whether preliminary results were consistent with hypotheses about multimodal instructions. The only modification made to the instrument was the addition of introductory text encouraging respondents to read and/or listen to definitions. Round 1 of data collection served as a first-phase pilot study, with 25 respondents in each of 4 treatment groups: control, spoken conventional, textual conventional, multimodal optimized. These three definition groups were selected because the unimodal conventional conditions were predicted to be less effective than their optimized counterparts (H2), while the multimodal optimized condition was predicted to be more effective than its conventional counterpart. That is, if multimodal treatments produced responses more consistent with definitions than unimodal treatments (H1), differences would be most detectable with these four conditions. This combination of treatment groups served as a proof-of-concept, because if this pilot data suggested no difference between multimodal optimized and unimodal conventional groups at this stage, it may have indicated that a full-scale collection under the current design was not warranted.

Round 1 included 100 respondents, round 2 included 60 respondents, and round 3 included 880 respondents, for a total of 1,040 respondents (see Table 3 for data collection schedule). For each round of data collection, separate requests for participants (MTurk “HITs”) were created for each treatment group and posted within minutes of each other. Any MTurk user who accepted a HIT was considered ineligible for all other portions of this study, that is, they could not accept another HIT in the same or another round of data collection. This combination of multiple HITs and restrictions on duplicate respondents effectively created random assignment per treatment group. This hypothesis is further explored below by comparing the demographic characteristics of respondents in each definition treatment group (see 2.5 Respondent demographics).

Table 3- Data collection schedule

	Round 1 (April 4, 2018)	Round 2 (June 18, 2018)	Round 3 (August 18-19, 2018)	Total
Control	25		80	105
Spoken conventional	25	30	160	215
Spoken optimized			160	160
Textual conventional	25		80	105
Textual optimized			80	80
Multimodal conventional			160	160
Multimodal optimized	25	30	160	215

Round 2 of data collection, the second-phase pilot study, included 30 respondents in each of the spoken conventional and multimodal optimized treatment groups. This round of data collection added instructions that specifically encouraged respondents to use definitions (“Please read/listen/both read and listen to all instructions carefully

before answering each question. This will help ensure that you provide the best information you can.”). Survey paradata confirmed that these instructions increased the percentage of respondents that played audio definitions and the percentage of definitions that were listened to (when compared to round 1), and those instructions were included in all 6 surveys with definitions during the round 3 collection. This was the only change made to the survey instruments between any rounds of data collection.

Round 3 of data collection served as the main data collection for this study. For each treatment without spoken definitions (control, textual conventional, textual optimized) data were collected from 80 respondents. In order to account for the fact that not all respondents listened to spoken definitions for every question, even with the added instructions, data were collected from 160 respondents for each treatment that included spoken definitions (spoken conventional, spoken optimized, multimodal conventional, multimodal optimized) to assure sufficient a sample size in these conditions when inclusion in the analyses required playing the audio definition.

Respondents were limited to MTurk users in the United States ages 18 or older.

Respondents were not able to complete more than one questionnaire, either within the same data collection round or across rounds. Each survey was given the same description within Mechanical Turk and posted within minutes of each other, effectively creating random assignment of respondents to treatment groups.

Respondents also were required to have an “approval” rate of 90% or higher for their Turk HITs, and to have completed at least 100 HITs.

Table 4 displays metrics about each definition component. For textual definitions, the Flesh-Kincaid reading level is shown as a measure of readability. For each question, the reading level for conventional definitions is higher than for optimized definitions. In addition, the estimated time needed for respondents to fully read the question text and associated definitions is shown, based on the word count of definitions and question text (more information below in 3.2 Compliance with textual definitions). Finally, this table shows the length of spoken definitions if audio files are played at a standard speed, with longer audio clips for spoken conventional than spoken optimized definitions. Optimized spoken definitions were shorter than conventional spoken definitions, and optimized textual definitions were shorter and less complicated than conventional textual definitions, supporting the theory that optimized definitions would lead to increased comprehension and data quality.

Table 4- Flesch-Kincaid reading level and estimated reading and listening time, by definition optimization, mode and question

	Conventional definitions			Optimized definitions		
	Textual		Spoken	Textual		Spoken
	Reading level	Estimated time to read (seconds)	Time to listen (seconds)	Reading level	Estimated time to read (seconds)	Time to listen (seconds)
1	9.8	15.2	25	2.8	8.6	12
2	11.4	17.2	26	7.4	9.2	12
3	11.4	14	20	15.6	5.8	13
4	9.3	16.8	29	6.9	8.6	17
5	10.7	15.2	24	6.4	11.4	16
6	12.8	19.2	32	4.9	11.8	20
7	8.3	11.2	18	3.5	6.8	9
8	9	12	23	7.6	7.6	10
9	10.1	11.6	20	5.8	4.4	8
10	10.9	13.6	22	6.9	7.8	10
11	10	12.8	25	4.5	7	14
12	6.8	12	19	5.9	6.2	11
13	10.8	12.2	17	3.9	8.2	7
14	10.8	15.4	26	9	6.6	6
15	12.5	12	20	7.7	5.4	6

2.4 Data cleaning

During data cleaning, data quality was evaluated at both the respondent- and item-level. If data from a given respondent suggested they did not attend to the survey or displayed otherwise suspicious behavior, all data from that respondent were removed. Next, responses to individual questions were examined, and implausible or outlying values were deleted while retaining all other valid data for those respondents.

First, data were examined at the respondent-level. Three open-ended questions asked respondents to list food items eaten the last 3 days that met particular criteria. When it was clear that respondents had not read the questions asked, for example, providing a definition of words included in the questions or a generic statement about food, the entire case was deleted. Cases that provided low-quality data but for which it was plausible the respondent had read the questions asked (for example, the respondent provided lists of food that did not meet the question criteria, numerical responses, or complaints about open-ended items) were not deleted. In total, 24 cases were deleted based on this criteria.

Despite parameters that respondents were only eligible to participate once in the study, in any round or treatment condition, 6 pairs of cases were identified as being completed by the same individual; members of these pairs used the same MTurk ID. For each of these individuals, the surveys were completed consecutively rather than concurrently, so the second case was removed from analysis.

In addition, 3 more responses were recorded than expected in the spoken optimized group, and 1 more response was recorded than expected in the multimodal conventional group, resulting in a total of 1,014 cases for analysis.

Next, implausible numeric responses were removed at the item-level. For example, questions 1 and 2 asked about hours of television and radio consumption during the past week. Given that there are 168 hours during a week, a response of 150 hours was deleted as logically implausible, though responses up to 100 were kept. Responses up to 100 hours were also kept in response to hours of internet, e-mail use, and work during the past week (questions 4, 5, and 7). For question 14 on the number of hours of exercise per week, responses up to 120 hours were kept, since they were feasible given the definition of “exercise” used in the survey.

For question 3 on the number of e-mails sent per week, implausible values of 400 and higher were deleted.

For question 8 on the number of miles traveled by ground vehicle within the past week, responses with the unlikely value of 6,000 or more was deleted. For question 9 on the number of plane trips within the past year, responses with the implausible value of 2,000 or more was deleted.

For question 10 on restaurant visits within the past month, responses with the implausible value of 346 or more was deleted. For question 11 on the number of pairs of shoes owned, responses with the unlikely values of 465 and higher were deleted.

For question 13 on the average hours of rest per day, responses above 18 were deleted, as well as responses of 3 or fewer hours. It should be noted that many, though not all, responses above 24 hours were also divisible by 7, perhaps indicating that respondents were using a reference period of one week rather than one day. All such responses were deleted, rather than assuming a weekly reference period was used and dividing large numbers by 7 days per week.

For question 15 on weekly caffeine consumption, responses over 140 were deleted as unlikely to be accurate responses.

The average responses provided to each question, after outlying data were removed, are provided in Appendix C- Mean response by question and definition mode and optimization.

In order to combine questions for analysis, responses were converted to a consistent scale, as the number of annual plane trips or weekly hours of television cannot be directly compared. Z-scores were computed for each response, using the average value and standard distribution of each survey item. In order to minimize the effect of

outliers, Z-scores were capped at a maximum absolute value of 4, since more extreme values would be unlikely to appear on a Z distribution

Finally, because some definitions were inclusive and expected to result in higher reported counts, while others were exclusive and expected to result in lower reported counts, Z scores for exclusive questions were multiplied by negative 1 so that regardless of the question type, higher Z scores indicated responses more aligned with definitions. A “long” data file was created with these Z scores, facilitating analyses in which individual responses or observations were the unit of analysis, rather than cases (respondents).

In total, 1,014 cases were used for analyses at the case level. For the 12 experimental survey questions, 11,988 individual responses were retained for analysis after removing implausible values and account for respondent-missing data. The distribution of respondents and observations by definition treatment is shown below in Table 5.

Table 5- Sample sizes by definition treatment

	Respondents	Observations
Control	104	1,239
Spoken conventional	200	2,356
Spoken optimized	162	1,920
Textual conventional	101	1,196
Textual optimized	80	952
Multimodal conventional	160	1,890
Multimodal optimized	207	2,435
Total	1,014	11,988

Note that the number of respondents and observations varies by definition treatment for two primary reasons. As shown in Table 3, the control, spoken conventional, textual conventional, and multimodal optimized groups were part of multiple rounds of data collection. In addition, more respondents were included in treatment groups with spoken components (both spoken and multimodal definitions) under the assumption that not all respondents would play spoken definitions.

2.5 Respondent demographics

During data collection, survey respondents were permitted to complete one questionnaire only. Each experimental treatment had the same description on Amazon Mechanical Turk, so respondents did not select which type of definitions they would prefer. Respondents were not assigned to treatment group based on their demographics, so their characteristics were examined during analysis to confirm whether the demographics of each treatment group were similar. Equivalence was expected given the study design. Any substantial differences between groups only become a concern if a characteristic is expected to be associated with analytic outcomes.

Overall, 55 percent of respondents were female and 44 percent were male; less than 1 percent identified as another gender. The percent of female respondents ranged from 43 percent in the control group to 64 percent in the multimodal optimized group, and a chi-square test indicated that the distribution of gender differed between each of the seven treatment groups ($\chi^2(12)=23.9684, p=.020$). While this indicates that the gender distribution respondents differed by treatment group, we have no reason to believe *a*

priori that response alignment with definitions would differ by respondent gender. While that relationship will be examined during analyses of respondent compliance and response alignment with definitions, it does not raise immediate concern.

Respondents ranged in age from 19 to 82, based on self-reported year of birth. The mean age was 30, and the median age was 39. While one-way ANOVA indicates that the mean age differed by treatment group ($F(6)=2.14$, $p=.0464$), this average ranged from 37 for the textual conventional group to 42 for the spoken optimized group. Although there is a statistical difference in age between treatment groups, the size of this difference does not seem likely to affect results. However, differences in compliance and response alignment with definitions will be examined in greater detail.

Overall, 73 percent of respondents described themselves as White and non-Hispanic. This percentage was similar across all 7 definition treatment groups ($F(6)=1.31$, $p=.2508$).

Respondents reported higher levels of education than is found in the general population (US Census Bureau 2019). About 12 percent had a high school diploma, its equivalent, or less as their highest degree. Another 39 percent reported having some college education but less than a bachelor's degree, 36 percent had a bachelor's degree, and 13 percent reported holding a degree above a bachelor's degree. This distribution was similar across all seven treatment groups ($\chi^2(18)=21.2871$, $p=.2652$).

Respondents were asked 4 questions about their use of technology. In particular, they were asked to report the frequency with which they send and receive text messages on a mobile phone, use apps on a mobile phone, watch television or movies on a computer, and search the internet for information on any device (never, several times a month, once a week, several times a week, once a day, several times a day or more). For each of these 4 items, chi-square tests revealed that there were no differences across experimental conditions for any of the technology use categories ($\chi^2(30)=27.2790, p=.6086$; $\chi^2(30)=30.7676, p=.4269$; $\chi^2(30)=34.0744, p=.2779$; $\chi^2(30)=17.6599, p=.9638$), and the overall distribution for each item is shown below in Table 6. A small amount of missing data was observed (2 cases for one item, 1 case for another). In addition, it should be noted that because all respondents were recruited from an online platform to complete a web survey, technology use is expected to be higher for study participants than for a general population.

Table 6- Distribution of responses by technology-based activity (How often do you...)

	Never	Several times a month	Once a week	Several times a week	Once a day	Several times a day or more	χ^2 p (df)
Send and receive text messages on a mobile phone	5%	6%	3%	20%	6%	60%	27.2790 .6086 (30)
Use apps (for any purpose) on a mobile phone	7%	7%	2%	67%	5%	13%	30.7676 .4269 (30)
Watch TV shows, movies, etc. on a computer	17%	14%	7%	20%	18%	23%	34.0744 .2779 (30)
Search the Internet for information on any device	<1%	7%	1%	79%	3%	9%	17.6599 .9638 (30)

To create a scale from these items, questions were combined such that response values 1 (never) through 6 (several times a day or more) were added for each respondent across questions, with a total possible value ranging from 4 to 24. Both the distribution of responses for individual items and the overall combined score were left skewed; about 12% of respondents selected the “several times a day or more” for all 4 items, resulting in the highest possible combined score. This combined value did not differ across treatment groups ($F(6)=0.89$, $p=.5030$). Because of the skewed nature of this summed combined measure, values were recoded into a high experience group (values of 22 and above), moderate experience (values of 17 to 21), and low

experience (values from 4 to 16). These represent the top quartile, middle two quartiles, and lowest quartiles of summed values, and this 3-point scale also did not differ across treatment group ($\chi^2(12)=12.6588, p=.3943$).

Other demographic characteristics may be related to the effectiveness of multimodal definitions. For example, this survey was administered in English only, and respondents were not asked whether English was their native language. Hearing loss among respondents is relevant when assessing spoken or multimodal definitions, and neither concept was measured in the survey. These factors are likely to be distributed evenly across treatment groups, but perhaps more relevant to some definitions types than others (for example, respondents with hearing loss may behave similarly to other respondents when presented with textual definitions, but not when presented with spoken definitions). Future research should consider whether these, and other unmeasured demographics, may be relevant to the experimental results.

Chapter 3: Compliance with definitions

This chapter reports empirical evidence about respondent compliance with spoken, textual, and multimodal definitions. Practitioners know that respondents often do not comply with survey instructions, and there is no reason to assume that will be different in this study. This chapter estimates respondent compliance with instructions to play spoken definitions and to read textual definitions, and compares compliance between conventional and optimized definitions and between unimodal and multimodal definitions. Self-reported reasons for compliance and lack of compliance are then discussed, followed by an analysis of respondent characteristics associated with compliance.

In the context of this study, compliance refers to whether respondents played or read definitions. Compliance is considered from the researcher's perspective, that is, whether respondents followed instructions. It is important to note that a respondent may provide high-quality responses without meeting these criteria, perhaps because the definitions are not applicable to their situation or because their interpretations of terms is consistent with the intent of the questions. The alignment of survey data with definitions is analyzed further in Chapter 4: Impact of definitions on data quality.

3.1 Compliance with spoken definitions

To measure respondents' exposure to spoken definitions, the online survey collected how many times a spoken definition was fully played. It is important to note that this measure could not record whether a respondent's audio was muted, nor whether they

truly attended to the spoken information, but instead serves as a proxy for respondent compliance in a self-interview setting. Audio clips were available to respondents in a standard media bar, and could be heard when respondents clicked the “play” icon. Playing spoken definitions could not be sped up in standard browsers, though a small number of respondents reported having sped up spoken definitions when asked about their compliance during debriefing questions. For unimodal spoken and multimodal definitions, a given response was considered “compliant” if the audio file was fully played.

For the four definition types with a spoken component (spoken conventional, spoken optimized, multimodal conventional, multimodal optimized), compliance with spoken definitions (that is, spoken definitions were fully played) ranged from the relatively low rate of 29 percent for multimodal conventional to 47 percent for spoken optimized (Table 7). The number of observations for which a definition was fully played at least once ranged from 553 for multimodal conventional to 932 for spoken conventional.

Table 7- Compliance with spoken definitions by question number and definition type

Question	Spoken conventional	Spoken optimized	Multimodal conventional	Multimodal optimized
1	73%	78%	61%	84%
2	59%	54%	40%	42%
3	41%	50%	32%	32%
4	44%	52%	30%	38%
5	39%	46%	27%	38%
6	36%	45%	26%	26%
7	26%	33%	24%	26%
8	34%	40%	22%	29%
9	37%	38%	26%	25%
10	40%	46%	20%	35%

11	25%	39%	21%	26%
12	31%	40%	21%	25%
Overall	39% (932)	47% (894)	29% (553)	35% (842)

For all four treatment groups, compliance was highest with the first definition presented in the survey, ranging from 61 percent for multimodal conventional to 84 percent for multimodal optimized. The compliance rate for the first survey question was significantly higher than the compliance rate for each subsequent question at the $p=.05$ level. This drop-off in compliance occurred both overall and within each of the 4 treatment groups with spoken definitions.

Overall, compliance rates were significantly different by treatment group ($F^1(3,725)=6.38, p=.0003$). In particular, compliance was higher for the spoken optimized group than for either of the multimodal conditions, and overall compliance was higher for the spoken conventional group than for multimodal conventional definitions. That is, for both conventional and optimized definitions, compliance with instructions to play the audio was higher for respondents who only received spoken definitions, rather than multimodal respondents who were encouraged to both read and listen to definitions. However, it should be noted that these overall compliance rates were less than 50 percent for each condition.

¹ Here, and for all F tests conducted at the observation level rather than the person level, multilevel models were used to account for potential within-respondent correlation, so the F-statistic refers to the results of a type 3 F test of fixed effects.

Differences in compliance between unimodal and multimodal groups may be driven by the presence of an alternative way of acquiring definition content with multimodal treatments. For respondents in unimodal spoken groups who were inclined to use definitions in their responses, their only choice was to listen to spoken definitions. Respondents in multimodal groups could have given responses consistent with definitions by reading textual definitions, even if they did not play an audio clip, and may have found the need to fully or partially listen to spoken definitions less pressing. Respondent compliance with each component of multimodal definitions is further discussed below (see 3.3 Compliance with multimodal definitions).

Comparing the overall compliance rates for optimized and conventional definitions, this rate was higher for respondents in optimized than conventional groups (47 and 39 percent, respectively, for unimodal; 35 and 29 percent, respectively, for multimodal), though this difference was only significant when comparing the two unimodal conditions. This pattern may be due to the nature of optimized definitions, designed to present key information in a succinct and conversational tone, but it cannot be disentangled from the relatively shorter duration of optimized definitions. As shown earlier in Table 4, each conventional spoken definition was longer than the optimized spoken definition for the same question. It is possible that respondents in the spoken conventional and spoken optimized groups both began to play spoken definitions and both spent the same amount of time on the question page before moving on to the next question, but only the spoken optimized respondent fully played the definition. In such a scenario, only the spoken optimized respondent would be considered

compliant for that survey item, even though the same amount of time may have elapsed. Part of the process of converting conventional definitions to optimized definitions involved shortening them, so length and optimization are fully confounded and cannot be disentangled in this study.

Compliance can also be explored by analyzing the number of questions for which each respondent fully played a definition. Below, Table 8 shows the distribution for how many definitions respondents listened to (0 through 12 definitions) by treatment group.

Table 8- Percentage distribution of number of questions for which a respondent fully played definitions by definition type

Number of questions	Spoken conventional	Spoken optimized	Multimodal conventional	Multimodal optimized
0	22%	19%	38%	14%
1	19%	17%	17%	34%
2	7%	4%	9%	9%
3	7%	6%	4%	7%
4	6%	2%	3%	4%
5	4%	7%	3%	2%
6	3%	1%	2%	4%
7	3%	3%	1%	<1%
8	7%	3%	4%	2%
9	5%	4%	4%	4%
10	2%	6%	3%	4%
11	4%	7%	4%	2%
12	15%	19%	10%	13%
Average	4.7 questions	5.5 questions	3.5 questions	4.1 questions

Only 14 percent of multimodal optimized respondents listened to no definitions, while 38 percent of multimodal conventional respondents listened to no definitions. Overall, respondents in the spoken optimized group listened to entire definitions for an average of 5.5 questions, followed by 4.7 questions for spoken conventional, 4.1

questions for multimodal optimized, and 3.5 questions for multimodal conventional. On average, for both unimodal and multimodal definitions, the average number of questions for which respondents complied with definitions was higher for optimized than conventional definitions. While there is certainly room for improvement for spoken definitions, this pattern of more compliant behavior with optimized than conventional definitions is a key design consideration, though more compliant behavior for unimodal definitions perhaps indicates that respondents may not prefer to play spoken definitions when presented with a text option.

3.2 Compliance with textual definitions

In order to estimate whether a respondent read the textual information presented to them, I estimated respondents' behavior using word counts and the amount of time spent on each page of the questionnaire. The word count for each question and, if applicable, textual definition was calculated. Instructions that appeared for each question prompting respondents to play a spoken definition were omitted from this word count. So, the word count for the control group and groups with unimodal spoken definitions were identical (question word count only), and the word count for the textual only and multimodal conditions were identical for conventional versions (question and conventional definition word count) and for optimized versions (question and optimized definition word count).

A threshold of 200 milliseconds (msec) per word was used to determine whether a respondent had sufficient time to read the text presented to them. This threshold was based on reading speed when learning textual information (Carver 1992). While this rate is more lenient than estimated reading speed in other survey research applications, for example, Zhang and Conrad (2014) assumed a reading speed of 300 msec per word, a slightly lower threshold was used here under the assumption that comprehension of complete thoughts (also described as "rauding" by Carver (1992)), and not necessarily longer-term retention, was a sufficient measure of compliance in the context of survey definitions that are not typically intended to be used beyond a question or battery of questions. The word count for each treatment and question was compared to the amount of time a respondent spent on a given page; respondents who spent at least as much time on the page as the product of 200 msec and the number of words presented were considered to have read the material. Because all respondents were shown written text (all survey items were text-based), this filter was calculated for respondents in all 7 treatment groups; respondents in the control and spoken

conditions were expected to have spent sufficient time on each page to have read question text. The time threshold for each question and conventional or optimized definition is shown in Table 4, and the time estimated to read optimized text was shorter than the time estimated to read conventional text for every definition.

However, like the proxy for spoken definition compliance, this criterion does not guarantee that respondents truly attended to and absorbed the textual information presented to them. Eye-tracking could help determine whether respondents viewed the textual information and analyze their eye movement patterns (for example, whether textual information is being viewed in order or if respondents are skipping or speeding through information), but even that would not capture whether they truly comprehended and internalized the information, or merely scanned the presented text. In a self-interview setting, time per page is the best proxy for respondent compliance.

For the four definition types with a textual component (textual conventional, textual optimized, multimodal conventional, multimodal optimized), Table 9 shows the percentage of respondents estimated to have read textual definitions.

Table 9- Compliance with textual questions and definitions by question number and definition type

Question	Textual conventional	Textual optimized	Multimodal conventional	Multimodal optimized
1	55%	88%	97%	99%
2	21%	40%	58%	72%
3	23%	76%	53%	84%
4	42%	61%	54%	74%
5	31%	66%	50%	82%
6	36%	70%	62%	82%
7	21%	64%	44%	84%
8	29%	64%	49%	74%
9	33%	74%	50%	74%
10	29%	59%	42%	67%
11	18%	48%	41%	67%
12	17%	70%	43%	78%
Overall	29% (352)	65% (617)	53% (1010)	78% (1905)

As shown in Table 9, estimated reading compliance was highest for the first survey item. Compliance was significantly higher for the first question than each subsequent question ($p < .0001$) for each definition type, similar to the pattern shown for compliance with spoken definitions. In addition, compliance rates differed by definitions treatment group for each pairwise comparison between the 4 definitions conditions with textual components. That is, the 78 percent compliance rate for multimodal optimized definitions was significantly higher than the 65 percent compliance rate for textual optimized definitions ($t(544) = 3.54$, $p = .0004$), which was

significantly higher in turn than the 53 percent compliance rate for multimodal conventional definitions ($t(544)=2.74, p=.0064$), which was significantly higher than the 29 percent compliance rate for textual conventional definitions ($t(544)=6.54, p<.0001$). So, compliance was highest for optimized definitions. For both conventional and optimized definitions, compliance was higher for multimodal than unimodal definitions. However, as with spoken definitions, all textual optimized definitions had fewer words than all textual conventional definitions, and therefore shorter estimated reading times. Length and optimization are fully confounded and cannot be disentangled in this study.

Table 10- Percentage distribution of number of questions for which text could be fully read by respondent and definition type

Number of questions	Textual conventional	Textual optimized	Multimodal conventional	Multimodal optimized
0	17%	0%	1%	0%
1	18%	8%	11%	2%
2	17%	4%	12%	<1%
3	10%	5%	13%	3%
4	8%	1%	8%	2%
5	5%	9%	3%	5%
6	7%	9%	9%	7%
7	5%	6%	1%	6%
8	4%	9%	8%	10%
9	3%	13%	3%	7%
10	2%	14%	5%	7%
11	3%	9%	8%	20%
12	2%	15%	19%	30%
Average	3.5 questions	7.7 questions	6.3 questions	9.2 questions

Table 10 shows the percentage of respondents in a given treatment group that are estimated to have read question and definition text. All respondents in the textual optimized and multimodal optimized groups were estimated to have read text for at least one question, and 99 percent of multimodal conventional respondents were estimated to have read text for at least one question. However, 17 percent of textual conventional respondents were estimated to have not read text for any questions. While respondents in both the textual and multimodal conventional groups were faced with the same, long, block of text to read, it is notable that their behavior differed. It is possible that the multimodal group, when given the option to listen to spoken definitions, spent more time on the page for several possible reasons including (1) the importance of definitions was emphasized by showing two communication modes, (2) they preferred to listen to the audio files partially or completely, or (3) they suspected from the presence of a media bar that their compliance or diligence would be tracked or tied to receipt of their promised incentive. No definitive

conclusion is possible from the main questionnaire, though debriefing questions asked of respondents who completed the main questionnaire may provide insight into this difference (see 3.6 Reasons for compliance).

Just 2 percent of respondents in the textual conventional condition were estimated to have read definitions for all 12 questions. Estimates of full compliance were 15 percent for textual optimized, 19 percent for multimodal conventional, and 30 percent for multimodal optimized. Overall, the average number of definitions estimated to have been read was higher for optimized definitions (9.2 for multimodal optimized and 7.7 for textual optimized) than for conventional definitions (6.3 for multimodal conventional and 3.5 for textual conventional).

3.3 Compliance with multimodal definitions

A respondent who reads the textual component but does not play the spoken component of a multimodal definition acquires information much like a respondent who reads a unimodal textual definition, and a respondent who only listens to the spoken component of a multimodal definition acquires information much like a respondent who listens to a unimodal spoken definition. Compliance with multimodal definitions depends on whether respondents only read, only listened to, or both read and listened to definitions. While Table 7 and Table 9 show the percentage of respondents that listened to a given definition or read a text definition, respectively, and Table 8 and Table 10 show the percentage of questions for which a respondent listened to definitions or read provided text, respectively, these do not fully describe respondent compliance when presented multimodal definitions. In order to understand respondent compliance, respondent behavior must be categorized into neither reading or listening to definitions, only listening to definitions, only reading definitions, or (for fully compliant respondents) both listening to and reading definitions.

Table 11- Compliance with multimodal definitions by question number and optimization

Question	Multimodal conventional				Multimodal optimized			
	None	Listened only	Read only	Listened and read	None	Listened only	Read only	Listened and read
1	3%	0%	35%	61%	1%	0%	15%	84%
2	41%	0%	18%	40%	28%	0%	30%	42%
3	48%	0%	21%	32%	16%	0%	52%	32%
4	46%	0%	24%	30%	26%	0%	36%	38%
5	50%	1%	23%	26%	18%	0%	54%	28%
6	38%	0%	36%	26%	18%	0%	56%	26%
7	56%	0%	19%	24%	16%	0%	58%	26%
8	51%	1%	28%	21%	26%	0%	45%	29%
9	49%	1%	25%	25%	26%	0%	48%	26%
10	58%	1%	22%	19%	32%	1%	33%	34%
11	59%	1%	20%	21%	33%	0%	41%	26%
12	57%	1%	22%	21%	22%	0%	53%	25%
Overall	46%	<1%	24%	29% (547)	22%	<1%	44%	35% (841)

For both conventional and optimized multimodal definitions, Table 11 shows the percentage of respondents who neither listened to nor read, only listened to, only read, or both listened to and read definitions by question. The percentage of respondents who are estimated to have only listened to a question is at or near 0 percent for most items, since for nearly all questions, the amount of time estimated to fully read the text on a given page is at least as long as the elapsed time of spoken definitions. That is, if a respondent fully played a spoken definition, enough time would have elapsed that they would be considered to have fully read the question and definition text.

For each question, the percentage of multimodal optimized respondents who both listened to and read the definition was at least as high as the percentage among respondents in the multimodal conventional condition. These differences ranged from

0 to 23 percentage points and overall, 35 percent of observations were both listened to and read for multimodal optimized respondents, compared to 29 percent for multimodal conventional. For each question, the percentage of multimodal conventional respondents who neither listened nor read was higher than the percentage among multimodal optimized respondents and overall, the percentage of observations that were neither read nor listened to was 46 percent for multimodal conventional respondents, compared to 22 percent for multimodal optimized.

Compliance with multimodal definitions can also be analyzed by respondent. In particular, Table 12 shows the number of questions in the multimodal conventional and multimodal optimized groups for which respondents were estimated to have both listened to and read definitions.

Table 12- Percentage distribution of number of questions for which definitions were listened to and questions and definitions could be fully read by respondent and definition type

Number of questions	Multimodal conventional	Multimodal optimized
0	38%	14%
1	17%	34%
2	9%	10%
3	4%	6%
4	3%	4%
5	3%	2%
6	2%	4%
7	1%	<1%
8	4%	2%
9	4%	4%
10	3%	4%
11	4%	2%
12	10%	13%
Average	3.4 questions	4.1 questions

About 38 percent of multimodal conventional respondents did not fully comply with any survey items, that is, there was no question for which they both read and listened to definitions, compared to just 14 percent of multimodal optimized respondents (Table 12). While 34 percent of multimodal optimized respondents complied for only one question compared to 17 percent of multimodal conventional respondents, the distributions of how many items were both listened to and read were otherwise similar. The average number of fully-compliant questions per multimodal conventional respondent was 3.4 questions and 10 percent listened to and read all items; the average number of fully-compliant questions per multimodal optimized respondent was 4.1 questions and 13 percent listened to and read all items. For each type of compliance metric, compliance with multimodal definitions appears to be higher for optimized than conventional presentations.

3.4 Self-reported compliance

After completing the substantive portion of the online survey, respondents were asked to report, if applicable, for how many questions they read definitions or listened to definitions (none, a few, most, all). For respondents in the spoken and multimodal treatment groups, additional questions were posed asking for how many questions they sped up playback of spoken definitions or muted their computer's sound. These responses are shown below in Table 13.

Table 13- Self-reported compliance with definitions by respondent and definition type

	Textual conventional	Textual optimized	Spoken conventional	Spoken optimized	Multimodal conventional	Multimodal optimized
READ DEFINITIONS	$\chi^2(9)=17.3159, p=.0435$					
None	2%	3%			1%	1%
A few	9%	4%			5%	4%
Most	23%	10%			24%	14%
All	66%	84%			70%	80%
LISTENED TO DEFINITIONS	$\chi^2(9)=29.2701, p=.0006$					
None			3%	7%	8%	5%
A few			26%	15%	36%	27%
Most			30%	29%	16%	29%
All			42%	49%	40%	38%
SPED UP PLAYBACK	$\chi^2(9)=12.4011, p=.1916$					
None			87%	94%	91%	87%
A few			10%	5%	5%	9%
Most			2%	0%	3%	3%
All			1%	1%	2%	1%
MUTED AUDIO	$\chi^2(9)=10.5394, p=.3086$					
None			97%	94%	93%	93%
A few			3%	3%	3%	3%
Most			1%	1%	4%	2%
All			0%	2%	1%	2%

Respondents appeared to overstate their compliance with all definition types. Sixty-six percent of textual conventional respondents reported they read all definitions, as well as 84 percent of textual optimized respondents. However, based on the number of words presented in questions and textual definitions, it was estimated that only 2 percent of textual conventional and 15 percent of textual optimized respondents had sufficient time to read all definitions (Table 10). For respondents shown unimodal

spoken definitions, 42 percent and 49 percent of those whose definitions were conventional and optimized, respectively, reported listening to all definitions (Table 13), even though audio files were fully played (based on tracking) for all questions by only 15 percent and 19 percent of respondents, respectively (Table 8). For respondents in the multimodal conventional condition, 70 percent reported reading and 40 percent reported listening to all definitions (Table 13). However, based on time spent per survey page and whether spoken definitions were fully played, only 19 percent of respondents read and 13 percent listened to all definitions (Table 8, Table 10). Similarly, for multimodal conventional respondents, 80 percent and 38 percent claimed to read and listen to all definitions (Table 13), while analysis indicated that only 10 percent and 13 percent read and listened to all definitions, respectively (Table 8, Table 10).

Few respondents indicated that they sped the playback of spoken definitions or muted their audio, and reports of this behavior were not different across treatment groups. These responses may be underestimates, since respondents who sped or muted definitions might not want to admit such actions. However, there is no objective measure (even an imperfect measure) to which to compare self-reporting adjustments to spoken definitions.

Time spent per page may overestimate respondent compliance with textual definitions for respondents who do not attend to the relevant text, but it may also underestimate respondent compliance for particularly quick readers. Tracking whether a spoken

definition was fully played may underestimate compliance by excluding observations where most, but not all, of the spoken definition was attended to (but appears unlikely to overestimate compliance given the relatively fewer number of respondents who reported muting their computer's audio). While acknowledging that the measures of respondent compliance used here are imperfect, it does appear that respondents overreported their compliance with both textual and spoken definitions.

3.5 Order of multimodal components

After completing the substantive portion of the online survey, respondents in the multimodal definition treatment groups who reported that they both read and listened to at least a few definitions were asked to best describe the order in which those activities occurred. Overall, 90 percent of respondents in the multimodal conventional group and 92 percent of respondents in the multimodal optimized group reported that they both read and listened to at least a few definitions (though, as discussed above, this most likely overestimates respondent compliance) and were asked whether they read definitions first, listened to definitions first, or did both simultaneously.

Table 14- Self-reported ordering of multimodal definitions

	Multimodal conventional	Multimodal optimized
I read before I listened	27%	42%
I listened before I read	13%	25%
I read and listened simultaneously	60%	33%

As shown in Table 14, self-reported order differed by the type of multimodal definition to which respondents were exposed ($\chi^2(2)=24.0077, p<.0001$). More multimodal conventional respondents reported that they both read and listened

simultaneously (60 percent) than either read first (27 percent) or listened first (13 percent). The self-reported order for multimodal optimized definitions was more evenly distributed between each option, with one-third reporting reading and listening simultaneously (33 percent), 42 percent reading first, and 25 percent listening first. The substantially higher reports of simultaneous listening and reading for the multimodal conventional group may be due to the longer duration of spoken conventional than optimized definitions; rather than waiting as long as 30 seconds for a definition to complete playing, respondents opted to read textual definitions while the audio files elapsed. Or, respondents who were shown multimodal optimized definitions may have been drawn to the short, clearly organized text on their screen before playing the audio file. Since respondents were not probed in detail about the order of their actions, and this order was self-reported rather than observed, both explanations are plausible but unknowable from these data.

3.6 Reasons for compliance

Respondents were asked to explain, in their own words, why they read textual definitions (or not) and why they listened to spoken definitions (or not). Follow-up questions were based on self-reported behavior, so, for example, if a respondent claimed to have listened to most spoken definitions, they were asked to explain why they listened to spoken definitions, even if it was later revealed that the respondent did not fully listen to any spoken definitions.

3.6.1 Textual definitions

Respondents in treatment groups with textual definitions (textual conventional and optimized, multimodal conventional and optimized) were asked whether they read all, most, a few, or none of those textual definitions. Respondents who read any (i.e., provided an answer other than “none”) were asked to describe why they chose to read definitions, and respondents who did not read all definitions (i.e., provided an answer other than “all”) were asked to describe why they chose not to read definitions. So, respondents who answered “all” were only asked why they read definitions, respondents who answered “none” were only asked why they did not read definitions, and respondents who answered “most” or “a few” were asked both why they read and did not read definitions.

Respondents shown unimodal textual definitions who reported reading those definitions explained that they attended to definitions as conscientious respondents.

They provided explanations such as:

- I wanted to know exactly what was being looked for.
- I wanted to get a clearer indication of what the researchers are looking for. I wanted to make sure I was doing the survey properly.
- I'm a "read the fine print" kind of person

Several noticed that definitions were particularly useful for answering questions. For example:

- You can't know the true meaning without the definition, which is why I chose to read them.
- The exclusions to definitions weren't always apparent.
- They informed me of information I didn't know I could use. For instance, I was allowed to include the time I spent commuting to work on the question that asked me how many hours I spent working in the past 7 days.

When unimodal text respondents did not read definitions, they did not believe definitions were necessary. These respondents “thought I knew what was wanted already” or “understood the basic gist of what you were looking for and did not feel the need to read further for certain tasks.”

3.6.2 Spoken definitions

Respondents in treatment groups with spoken definitions (spoken conventional and optimized, multimodal conventional and optimized) were asked whether they listened to all, most, a few, or none of those spoken definitions. Respondents who listened to any (i.e., provided an answer other than “none”) were asked to describe why they chose to listen to definitions, and respondents who did not listen to all definitions (i.e., provided an answer other than “all”) were asked to describe why they chose not to read definitions. So, respondents who answered “all” were only asked why they listened to definitions, respondents who answered “none” were only asked why they did not listen to definitions, and respondents who answered “most” or “a few” were asked both why they listened to and did not listen to definitions.

Some respondents, when presented with unimodal spoken definitions, suspected that the survey might include attention checks or other metrics that could affect their payment for completing the survey, motivating them to listen. Their reasons for playing spoken definitions include:

- to see if they had me enter a code
- I clicked play automatically at first, thinking it might be an attention check type thing. Then I kept clicking while I thought about the questions.
- Because the instructions said so and I did not want to get penalized for the HIT.
- I thought that I had to or the HIT wouldn't be approved

Others explained they were simply following instructions (“I was told to” or “Because the instructions said to listen to them”). Respondents also listened in order to provide quality data and be sure their responses aligned with the questions posed:

- They explained exactly what was meant by the question.
- To clarify my understanding of the parameters of the question.
- I wanted to answer the questions correctly. I needed to know what you wanted, so it was easy to listen to the definitions.
- I wasn't really going to but after listening to the first few, I noticed that we were supposed to exclude some things so I thought it would be best to listen to the rest

Some respondents explained that the definitions were useful and provided some surprising information that did change their interpretation of survey questions:

- I wanted to hear the specific definition to answer the question thoroughly.
Some of the questions had more specific definitions than I had initially expected, so listening to them did alter my questions to be more exact.
- In case there was information I should know, such as in the case of listening to the radio. If I wouldn't have listened, I would have thought you also wanted this to include music, but you didn't.
- I needed to know what they consider an acceptable answer. For instance, I watch 6-8 hours of TV a day but it is all through a streaming service so the answer to the TV question was "0".

When unimodal respondents did not listen, they typically believed definitions were not crucial or not worth the effort ("it got annoying"):

- The questions seemed self explanatory
- I stopped listening to the definitions because they had been all exactly what I thought the question was asking for. So it no longer seemed like a good idea.

Some respondents did not want to take the time to play spoken definitions. Two spoken optimized respondents explained they did not listen "to save time" and "because some were long." While spoken optimized definitions were shorter than spoken conventional definitions, even optimized clips may have felt too burdensome for respondents who wanted to quickly complete the survey.

Two respondents noted they did not play definitions because they noise may bother others near them (“sitting in room with others” and “I’m working on MTurk while my baby is sleeping and I didn’t want to wake him”), while another was unable to play audio files in their browser (“I am working in safe mode and cannot play sound”).

3.6.3 Multimodal definitions

Multimodal respondents who both read and listened to definitions provided two types of explanations for their compliance. Some respondents both read and listened in order to increase their understanding of the survey definitions provided. For example:

- Reading and listening at the same time helps my reading comprehension. It is easier and takes less work. I also liked listening to the reader’s voice. It was oddly soothing.
- Audio prompts are much better when paired with visual, so i think i naturally just started to play the audio in order to hear it as i read
- Because it helped with the comprehension of the written instructions.
- I wanted to read the directions first so I could comprehend them and then have them explained more thoroughly through the audio... I wanted to have them explained more thoroughly than just by reading them.

Like unimodal spoken respondents, some multimodal respondents explained that they complied to meet the requirements of the task (“You put up an audio file. Therefore you wanted me to listen to it. Again this is to give you the best product I can.”) or

because they suspected verification may be involved (“To make sure there were no contradictions from what was spoken.”)

While tracking indicated that few respondents only listened to spoken definitions, two respondents described that it was easier for them to listen to definitions than read them (“I was lazy to read it...[read] from top to bottom once I was done listening to the explanation”).

Multimodal respondents who primarily read explained that it was easier, faster, or more convenient with both reading and listening to definitions. For example, “I read far faster than the speed at which the definitions were read and I didn't want to wait when I could easily read.” Respondents also noted that persistent definitions are easier to reread or reference than spoken information, for example, “It is easier to go back over something in writing than it is to keep rewinding a recording.”

3.7 Predictors of compliance

Compliance with definitions is expected to result in responses more consistent with definitions (analyzed further in Chapter 4: Impact of definitions on data quality), so it is important to understand the factors associated with respondents' compliance with definitions. While most demographic characteristics were balanced across different treatment groups, gender and age showed some variation across the seven experimental groups (see 2.5 Respondent demographics), so it is important to determine whether, if definition alignment does differ across definition types, those

differences could be explained by differences in respondent demographics rather than definition mode and optimization.

For this analysis, questions (denoted with subscript q) are nested within respondents (denoted with subscript i). Both questions and respondents are given random intercepts, allowing for baseline differences in question difficulty and respondents' behavior, though respondents are considered to be random effects and questions to be fixed effects.

$$\log \left[\frac{C_{iq}}{1 - C_{iq}} \right] = \gamma_{00} + \gamma_{10}(G_i) + U_{i0} + e_{iq}$$

The dependent variable is compliance for a given observation, as predicted by definition treatment group G, a categorical variable with 7 different values for each definition type (control, conventional and optimized spoken, conventional and optimized textual, conventional and optimized multimodal). Additional predictors of respondent demographics are added for the characteristics discussed below, as well as an interaction term between demographics and treatment group. Compliance is a binary outcome, so logistic regression is used.

3.7.1 Gender

While the distribution of respondent gender differed across the 7 definitions treatment groups, gender did not predict compliance with definitions ($F(2,1005)=0.25$, $p=.7760$). This lack of significance held across definition types; when an interaction between gender and treatment was added to the model, it was not significant

($F(7,998)=1.88, p=.0697$), and gender continued to not be a significant predictor of compliance ($F(2,998)=0.92, p=.3982$).

3.7.2 Age

Respondent age was a significant predictor of compliance ($t(1006)=50.59, p<.0001$), with older respondents more likely to comply. However, this estimated effect corresponds to an odds ratio of just 1.0057. So, while age should be considered in further analyses (particularly to determine whether age is related to the effectiveness of definitions), this difference does not raise immediate concerns. When an interaction between age and definition type was added, it was not significant ($F(6,1000)=1.60, p=.1431$), though age remained statistically significant. That is, older respondents' higher level of compliance held across all types of definitions. This may be due in part to the criteria used to determine compliance with spoken definitions. An identical reading speed is assumed for all respondents, regardless of their age and cognition. However, older respondents may read more slowly than younger respondents (Liu, Patel, and Kwon 2017), so, for example, if an older and younger respondent both read only a portion of the provided text, it is possible that the older respondent would be, mistakenly, categorized as compliant while the younger respondents would be categorized, correctly, as noncompliant.

3.7.3 Race and ethnicity

Respondent race and ethnicity was not related to compliance with definitions. In particular, White, non-Hispanic respondents had similar compliance rates as non-White or Hispanic respondents ($t(1006)=0.00, p=.9713$). However, this effect did

vary by the type of definition available to respondents. Overall, there was a significant interaction between race/ethnicity and definition type ($F(6,1000)=2.63$, $p=.0155$). In particular, for textual conventional and spoken conventional definitions, White, non-Hispanic respondents were less likely to comply with definitions ($t(99)=2.38$, $p=.0193$ for textual conventional, $t(198)=2.04$, $p=.0431$ for spoken conventional). However, for multimodal optimized definitions, White, non-Hispanic respondents were more likely to comply ($t(205)=2.01$, $p=.0463$).

In addition, race/ethnicity becomes significant as a main effect ($t(1000)=2.11$, $p=.0350$), with White, non-Hispanic respondents complying at an odds ratio of 1.1176 compared to non-White or Hispanic respondents. However, the overall effect of race/ethnicity, when combined with an interaction term, was not significant ($t(1000)=0.02$, $p=.8988$). These mixed findings suggest that race may be an important consideration, and while differences generally indicate that non-White or Hispanic respondents are more compliant with definitions, this pattern may not generalize.

3.7.4 Education

Compliance with definitions varied by respondent education. Overall, education was a significant predictor of compliance ($F(3,1003)=2.76$, $p=.0413$). In particular, compliance was higher for respondents with a master's degree or higher than those with a high school diploma or less ($t(1003)=2.39$, $p=.0170$), corresponding to an odds ratio of 1.1069, and those whose highest degree was a BA ($t(1003)=2.13$, $p=.0331$), corresponding to an odds ratio of 1.0761. However, when an interaction between education and definition type was added, neither education ($F(3,985)=1.83$, $p=.1403$)

nor the interaction between education and definition type ($F(18,985)=1.29, p=.1874$) was significant.

3.7.5 Technology use

As introduced in 2.5 Respondent demographics, a variable capturing respondents' experience and familiarity with technology was computed. Observations from respondents with lower levels of technology experience were associated with higher levels of compliance; observations from the most experienced technology group were less compliant with definitions than observations from those with moderate ($t(1002)=2.11, p=.0347$) and lower technology experience ($t(1002)=2.89, p=.0039$). There was no significant interaction between technology experience and definition type ($F(12,990)=.83, p=.6147$). While it was initially expected that respondents with higher levels of technology experience may be most compliant with definitions, particularly due to their extensive internet experience, this finding may make sense in the context of the experiment. Respondents with relatively low levels of technology experience may be more motivated to show compliance with the tasks in front of them, or respondents with the highest levels may be confident in their ability to complete the tasks without additionally reading or listening to the definitions presented. However, it should be noted that the relatively low level of technology experience is relative to an extremely savvy group of technology users, after all, respondents are all users of an online crowdsourcing platform and the scale used to measure the experience may have been better suited to a more heterogeneous population. Nonetheless, this pattern is certainly notable and could be further

explored in an experiment that uses a different recruitment method or sample that is more reflective of the general population's relationship with technology.

3.8 Summary

Respondents did not fully comply with unimodal or multimodal definition instructions, though this pattern is certainly consistent with past research on respondent compliance with survey instructions (Conrad et al. 2006; Peytchev et al. 2010). Compliance with spoken definitions was of particular interest, since this represents a new application of spoken communication in a primarily visual mode. Compliance with spoken definitions was highest when respondents were presented with unimodal spoken optimized definitions, for which nearly half of definitions were fully played, although all compliance rates show room for improvement.

For each type of definition (unimodal spoken, unimodal textual, multimodal), compliance was higher for optimized than conventional definitions. Optimized definitions were designed to be more appealing to respondents, so it is not surprising that compliance was higher, either because respondents preferred their layout or the relatively shorter duration needed to read or listen to them. With this study design, it is not possible to determine whether the shorter duration or the techniques used to optimize definitions for the mode in which they were presented account for this finding, since, shortening definitions was part of the optimization process

For spoken definitions, compliance was higher for unimodal than multimodal presentations, while for textual definitions, compliance was higher for multimodal

than unimodal presentations. This may indicate respondents' preference for textual information when presented with multimodal definitions. The higher compliance rates when adding a spoken component to a textual definition (that is, comparing multimodal and unimodal textual), may not be due solely to the novel appearance of an audio clip in a visual survey, since those differences persisted throughout the entire questionnaire.

For all types of definitions, compliance was highest for the first survey item, then lower for subsequent items. Respondents were open to all definition types initially, but relatively few met compliance criteria for all survey items. While some respondents reported that they found the definitions helpful or particularly liked spoken or multimodal formats, others indicated that definitions were not useful to them or involved more time and effort than they would prefer to expend.

Some demographic characteristics appear to be related to respondent compliance with definitions. Surprisingly, observations from respondents with less technology use were more compliant with definitions than observations from respondents with the highest levels technology use, though this finding may be unique to the relatively high level of technology among the respondent sample (Amazon Mechanical Turk workers). All of these demographic relationships, particularly technology use, should be evaluated for other populations of interest and for other types of multimodal communication. However, these results show that for this population, respondents are

willing to comply with instructions to listen to spoken definitions, either in unimodal or multimodal presentations.

Chapter 4: Impact of definitions on data quality

This chapter evaluates the effectiveness of definition type by comparing the degree to which survey responses were aligned with definitions by the type of definition shown to respondents, in order to assess the effectiveness of definition mode (unimodal textual, unimodal spoken, or multimodal) and optimization (conventional or optimized). Chapter 3: Compliance with definitions explored the extent to which respondents complied with instructions about how to use each type of definition. In this chapter, the effectiveness of definitions is evaluated for all observations, followed by a closer analysis of the relationship between compliance and definition effectiveness. Finally, respondent characteristics are analyzed to determine whether the effectiveness of different types of definitions depends on respondent demographics.

4.1 Definition mode

This section compares responses given to unimodal textual, unimodal spoken, and multimodal definitions. Responses are first examined overall, that is, regardless of whether they were compliant with definitions. Then, compliance is considered, both by analyzing the subset of observations that were generated in compliance with instructions about definitions use, and by comparing results by the treatment effectively received when noncompliance for particular observations is taken into account. Only mode, and not optimization, is considered in this section, so each mode group includes both conventional and optimized definitions (for example, the

unimodal textual group includes both conventional text and optimized text definitions).

As described in 2.3 Data collection, survey questions used different reference periods and asked about different types of activities. Responses could not be averaged as provided, both because each question operated on a different scale and because for some questions higher numeric responses indicated compliance with definitions, and for other questions lower numeric responses indicated compliance with definitions. Instead, the responses for each question were converted to a Z-score, trimmed to +4 and -4, and these Z-scores were multiplied by -1 for questions with exclusive definitions. This conversion allowed responses to be pooled across questions for analysis, using a consistent scale for each item (Z-scores) and having a consistent interpretation of these Z-scores such that higher values indicate values more aligned with definitions, while lower values indicate values less aligned with definitions.

4.1.1 Overall alignment with definitions

First, average Z-scores were compared for each type of definition to which a respondent may have been assigned: unimodal textual, unimodal spoken, multimodal, and the control group (no definitions). As discussed in 2.4 Data cleaning, extreme values for Z-scores were trimmed to +4 and -4 in order to allow for variation in responses while minimizing the impact of outlying values.

While a multilevel model was used to account for potential correlation within respondents, multiple measures per respondent explained very little of the variation in responses. Less than 1 percent of variation was due to within-respondent differences (variance of 0.002946 within respondents and 0.7713 for residual variance). Table 15 shows the average Z-score by definition mode (and a more detailed presentation by question can be seen in Appendix B- Mean Z-score by question and definition mode and optimization). Higher values indicate more alignment with definitions, while lower values indicate less alignment with definitions. The mean of all responses is 0, so Z-scores indicate the number of standard deviations by which observations for a given definition mode varied from the average response.

Table 15- Mean Z-score by definition mode

Definition mode	Mean Z-score
Control (no definition)	-0.1260
Spoken	-.00854
Textual	-.01405
Multimodal	0.0410

Responses to questions with multimodal definitions were more aligned with definitions than the average response by about 0.041 standard deviations. Responses to questions with unimodal spoken and textual definitions were less aligned with definitions than the average response by about 0.008 and 0.141 standard deviations, respectively, though all definition types yielded responses that were more aligned with definitions than the control group that was not exposed to definitions (less aligned with definitions than the average response by about 0.126 standard deviations).

A general linear mixed model was used to compare the effects of different definition treatments by modeling Z-scores while accounting for clustering of observations within respondents. In particular, Z-scores were modeled while accounting for up to 12 observations for each respondent. Because not all respondents provided values for all 12 experimental questions (and some respondents provided implausible values that were deleted during cleaning), a model that would result in casewise deletion, such as a repeated-measures ANOVA, was eliminated. With a mixed model, if an observation is missing for a respondent, that respondent's otherwise valid data can be included in analyses.

For this analysis, questions are nested within respondents. Respondents were given random rather than fixed intercepts, in order to allow for baseline differences in behavior. Questions were given fixed intercepts, since Z-scores were calculated using for each question. Because Z-scores were trimmed, the average value for each question may deviate slightly from 0. Respondents were considered to be random effects and questions to be fixed effects.

Responses to multimodal definitions were significantly more aligned with definitions than each other definition type. That is, average Z-scores were higher for the multimodal group than unimodal textual definitions ($t(996)=2.33, p=.0203$), unimodal spoken definitions ($t(1002)=2.56, p=.0105$), and the control treatment with no definitions ($t(989)=5.77, p<.0001$). Overall, definition mode was a significant predictor of the degree to which responses were aligned with definitions

($F(3,994)=11.37, p<.0001$). There was no significant difference between responses provided to textual and spoken definitions.

4.1.2 Alignment as a function of compliance

Respondents were not fully compliant with instructions for attending to definitions, raising the possibility that the effect of definition type on alignment might differ as a function of compliance. To address this, observations can be compared both overall and by examining only observations estimated to be in compliance with the experimental treatment. For spoken definitions, we captured whether or not audio clips were fully played, and for textual definitions, we estimated the time that would be required to read a particular question and its associated definition text and compared this to the actual time each respondent spent on the page. More detail about these measures is presented in sections 3.1 Compliance with spoken definitions and 3.2 Compliance with textual definitions. The control group with no definitions is included in these analysis, since respondents were expected to have spent sufficient time on each page to read survey questions.

The average Z-score by definition mode for compliant observations is shown below in Table 16. As with Table 15, Z-scores show the number of standard deviations by which observations for a given definition mode differ from the average response, with higher values indicating more alignment and lower values indicating less alignment.

Table 16- Mean Z-score by definition mode for compliant observations

Definition mode	Mean Z-score
Control (no definition)	-0.12878
Spoken	.07601
Textual	.05865
Multimodal	.16298

Responses to questions with multimodal definitions were more aligned with definitions than the average response by about 0.16 standard deviations. Responses to questions with unimodal spoken and textual definitions were more aligned with definitions than the average response by about 0.08 and 0.06 standard deviations, respectively.

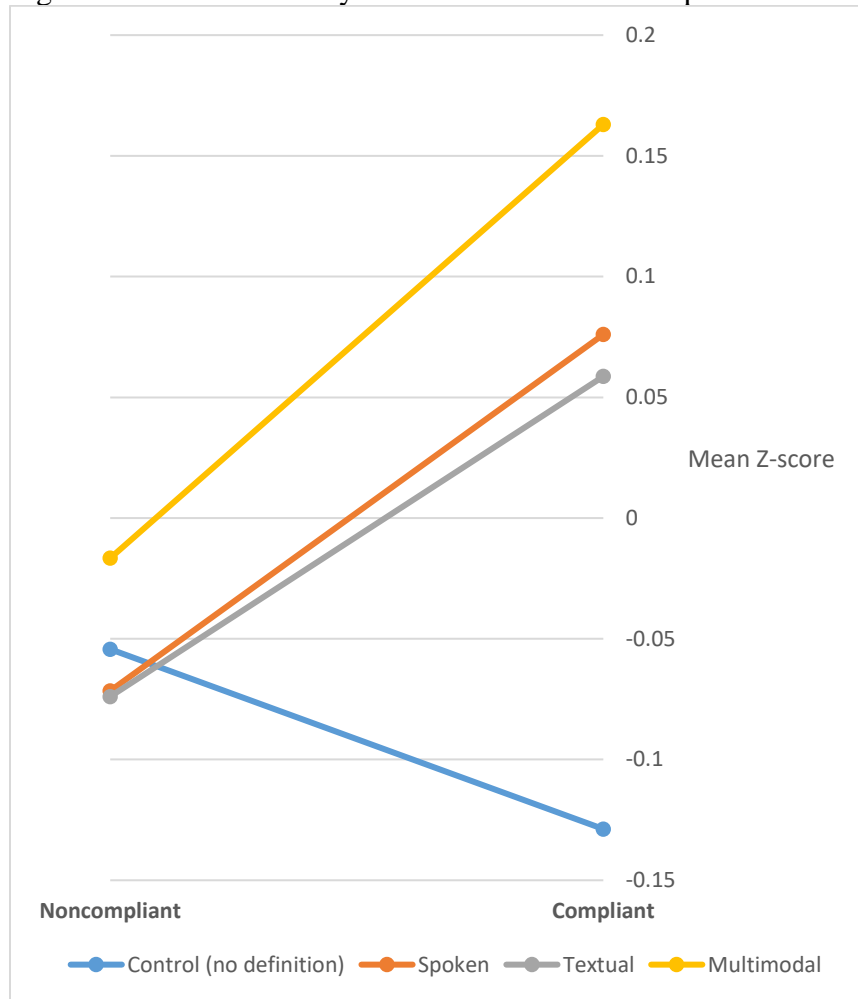
When multimodal definitions were presented and observations were compliant with definitions, those observations were significantly more aligned with definitions than each other definition type. That is, average Z-scores were higher for definitions that were multimodal than unimodal textual ($t(691)=2.86$, $p=.0044$), unimodal spoken ($t(619)=2.78$, $p=.0056$), or the control treatment with no definitions ($t(466)=8.32$, $p<.0001$). Overall, definition mode was a significant predictor of the degree to which responses were aligned with definitions ($F(3,552)=24.05$, $p<.0001$). There was no significant difference between responses provided to textual and spoken definitions.

Finally, the impact of definition compliance can be evaluated by using an interaction term (rather than just evaluating the subset of compliant observations) to determine whether compliance plays a different role across different definition modes, and to determine whether compliance fully explains these differences (rendering mode irrelevant). While mode is not significant overall when the model includes a term for

the interaction between compliance and mode ($F(3,2048)=2.16, p=.0905$), multimodal definitions yield responses that are significantly different from unimodal spoken definitions ($t(1369)=2.27, p=.0233$). The difference between responses given to multimodal and unimodal textual definitions was not statistically significant using a threshold of $p=.05$ ($t(1615)=1.89, p=.0586$).

The interaction between compliance and definition mode was significant overall ($F(4,3396)=20.06, p<.0001$), suggesting that the effects of mode were amplified for compliant observations. When definitions were presented, compliant observations were significantly more consistent with definitions for spoken ($t(2466)=5.46, p<.0001$), textual ($t(4552)=3.55, p=.0004$), and multimodal definitions ($t(2400)=6.32, p<.0001$). The average Z-scores for compliant and noncompliant responses by definition mode are shown in Figure 1.

Figure 1- Mean Z-score by definition mode and compliance status



The magnitude of compliance effects by definition mode can be seen in Figure 1. As noted above, the impact of compliance was not statistically significant for the control group, so the apparent decrease in response alignment with definition is not cause for concern. For textual and spoken definitions, compliant observations were about 0.15 and 0.13 standard deviations higher, respectively, than their noncompliant counterparts. For multimodal definitions, compliant observations were about 0.18 standard deviations higher than noncompliant responses. As noted above, the increase in definition alignment due to compliance was statistically significant for each definition type: textual, spoken and multimodal.

4.1.3 Intent to treat

The effectiveness of definitions can also be examined by focusing on the de facto treatments observations effectively received, rather than the treatment to which respondents were assigned. The previous analyses group observations by their intended treatment, then examine the subset of compliant observations and the effect of compliance within assigned groups. This section estimates the definition treatment observations received and analyzes those groups, in some cases, allowing respondents to behave as if they were in a different treatment than the one to which they were assigned (within the parameters of the study design).

For any observations for which a respondent in a spoken condition did not play the definition, they received a similar treatment as a control group participant and can be analyzed with that group. Since only whether a spoken definition was fully played was captured in this data collection, rather than how much was played, observations for which a spoken definition was partially played are considered noncompliant, so it should be noted that the treatments received are not necessarily identical, as some of the key definition information may have been conveyed for partially-compliant observations. For any observations for which a respondent in textual condition did not read the definition, they received the same treatment as a control group participant and can be analyzed with that group, though again, respondents may have read part of a definition or read all text more quickly than the target reading speed, so their experiences may not be identical. For any observations for which a respondent in a

multimodal condition did not both fully listen to and read the definition, they can be analyzed with the control, unimodal textual or (unlikely given the relative length of audio clips and compliance threshold for textual definitions) unimodal spoken groups. As with spoken definitions, observations were categorized based on whether they fully met compliance criteria, so observations for which definitions were partially played or read were considered noncompliant and analyzed accordingly.

Because compliance is measured for each observation in each treatment, this analysis by “treatment received” is also conducted at the observation level. For example, a multimodal respondent could have some of their responses categorized as multimodal, unimodal textual, and control, depending on their behavior across questions. Note that observations were excluded when they did not meet the minimum criteria for the control group, that is, so little time was spent on a given survey page that the question could not have been read at the assumed reading speed for respondents. The average Z-score for treatment effectively received is presented in Table 17.

Table 17- Mean Z-score by definition mode received

Definition mode	Mean Z-score
Control (no definition)	-0.07063
Spoken	.07481
Textual	.07575
Multimodal	.16298

For observations that complied with the multimodal treatments to which they were assigned², responses were more aligned with definitions by about 0.16 standard deviations than the average response. Responses that complied with instructions for spoken definitions, whether they were assigned to that group or a multimodal treatment, were more aligned with definitions by about 0.07 standard deviations than the average response, and responses that complied with instructions for textual definitions, whether they were assigned to that group or a multimodal treatment, were more aligned with definitions by about 0.08 standard deviations than the average response. Responses that received the control condition (because they were assigned to the control group and met the time threshold for reading question text, assigned to a unimodal definition treatment group and did not meet compliance criteria, or were assigned to a multimodal definition treatment group and did not meet the compliance criteria for either the spoken or textual components) were less aligned with definitions by about 0.07 standard deviations than the average response.

Observations compliant with multimodal definitions were significantly more aligned with definitions than each other definition type. That is, average Z-scores were higher for the multimodal group than when the effective treatment was textual definitions ($t(2567)=2.89, p=.0038$), spoken definitions ($t(1517)=2.88, p=.0040$), and the control treatment with no definitions ($t(1713)=8.99, p<.0001$). Overall, effective definition

² Observations from multimodal definitions that both read and listened to definitions were considered fully compliant. The average Z-scores for this group are identical whether noncompliant observations are removed from analysis, such as in Table 16, or grouped with another definition type, such as in Table 17.

mode was a significant predictor of the degree to which responses were aligned with definitions ($F(3,2057)=37.57, p<.0001$). There was no significant difference between responses provided to textual and spoken definitions ($t(1874)=0.01, p=.9903$).

For observations for which the effective treatment received was unimodal textual, the difference between responses that were assigned to a unimodal treatment and complied and responses that were assigned to a multimodal treatment group and did not comply was not statistically significant. That is, compliant unimodal and noncompliant multimodal responses were similar for textual ($t(259)=0.63, p=.5320$) definitions ($t(473)=0.83, p=.4071$).

4.2 Optimization

While the section above compares responses by definition mode (unimodal textual, unimodal spoken, or multimodal), this section further explores mode differences by comparing conventional and optimized definitions within each mode. That is, while multimodal definitions yield responses more consistent with definitions than unimodal definitions, those comparisons collapse conventional and optimized definitions into a single group. Optimized definitions were predicted to yield responses more consistent with the underlying concepts than conventional definitions, and such collapsing may conceal differences between unimodal conventional and optimized definitions. Further analysis may also help explain why multimodal definitions proved effective; if information is conveyed through independent channels (such as visual and auditory communication, as discussed in 1.3.1 Working memory),

then any multimodal communication is expected to outperform unimodal communication. However, if redundant information conveyed through multiple channels leads to a decrease in learning and comprehension (see 1.3.2 Redundancy effect), then only multimodal optimized definitions are expected to outperform unimodal definitions.

This section compares responses given to each of the 7 treatment groups: spoken conventional, spoken optimized, textual conventional, textual optimized, multimodal spoken, multimodal optimized, and the control group without definitions. As with the section above, observations are examined both overall and for the subset of compliant observations. Finally, observations are grouped by the treatment they effectively received by reclassifying noncompliant observations based on behavior, rather than assigned treatment groups. The outcome in all analyses is Z-scores, as introduced in 4.1 Definition mode. Higher values indicate more alignment with definitions, lower values indicate less alignment with definitions, and 0 is the average alignment with definitions across all observations.

4.2.1 Overall alignment with definitions

As with previous analyses, multilevel models were used to account for potential correlation within respondents. In practice, multiple measures per respondent explained very little of the variation in responses. Less than 1 percent of variation was due to within-respondent differences (variance of 0.002740 within respondents and 0.7713 for residual variance). Using multilevel models, Table 18 shows the average Z-score by definition mode.

Table 18- Mean Z-score by definition mode and optimization

Definition mode	Mean Z-score
Control (no definition)	-0.12602
Spoken conventional	-0.02497
Spoken optimized	0.01163
Textual conventional	-0.03164
Textual optimized	0.00804
Multimodal conventional	0.01309
Multimodal optimized	0.06272

Responses to questions with multimodal optimized definitions were more aligned with definitions than the average observation by about 0.0627 standard deviations. That is, average Z-scores were higher for the multimodal optimized group than for spoken conventional ($t(1003)=3.39, p=.0007$) and textual conventional definitions ($t(997)=2.98, p=.0029$). Multimodal optimized Z-scores were higher, though not significantly so when using a threshold of $p=.05$, when compared to spoken optimized definitions ($t(997)=1.87, p=.0624$), textual optimized definitions ($t(990)=1.59, p=.1111$), or multimodal conventional definitions ($t(1000)=1.80, p=.0717$). That is, observations for multimodal optimized definitions were significantly more aligned with definitions than observations for unimodal conventional definitions, although they were not significantly different from unimodal optimized and multimodal conventional definitions. Overall, definition treatment group was a significant predictor of the degree to which responses were aligned with definitions ($F(6,993)=6.72, p<.0001$).

For unimodal definitions, optimization did not lead to a difference in responses. For unimodal spoken definitions, the differences between optimized and conventional definitions was not statistically significant ($t(996)=1.33, p=.1832$), though this

difference was in the predicted direction with higher Z-scores for optimized definitions. The same was true for unimodal textual definitions; the difference between textual optimized and textual conventional definitions was not statistically significant ($t(988)=1.02$, $p=.3081$) but showed more definition alignment with optimized definitions.

4.2.2 Alignment as a function of compliance

Respondents were not fully compliant with instructions for attending to definitions, and observations can be compared both overall and by examining only observations estimated to be in compliance with the study treatment. More detail about these measures is presented in sections 3.1 Compliance with spoken definitions and 3.2 Compliance with textual definitions.

The average Z-score by definition mode for compliant observations is shown below in Table 19. As with Table 18, Z-scores show the number of standard deviations by which observations for a given definition mode differ from the average response, with higher values indicating more alignment and lower values indicating less alignment

Table 19- Mean Z-score by definition mode and optimization for compliant observations

Definition mode	Mean Z-score
Control (no definition)	-.12878
Spoken conventional	.07291
Spoken optimized	.07880
Textual conventional	.01818
Textual optimized	.08181
Multimodal conventional	.07927
Multimodal optimized	.21742

Responses to questions with multimodal optimized definitions were more aligned with the underlying concepts than the average response by about 0.22 standard deviations. For compliant person-items, multimodal optimized definitions led to responses that were more aligned with the underlying concepts than every other definition treatment group. That is, average Z-scores were higher for the multi-modal optimized group than for the spoken conventional ($t(649)=3.41, p=.0007$), spoken optimized ($t(601)=3.26, p=.0012$), textual conventional ($t(959)=3.46, p=.0006$), textual optimized ($t(614)=2.97, p=.0006$), and multimodal conventional definition ($t(643)=2.78, p=.0056$) groups, as well as the control group with no definitions ($t(508)=8.65, p<.0001$). Overall, definition mode was a significant predictor of the degree to which responses were aligned with definitions ($F(6,600)=13.53, p<.0001$).

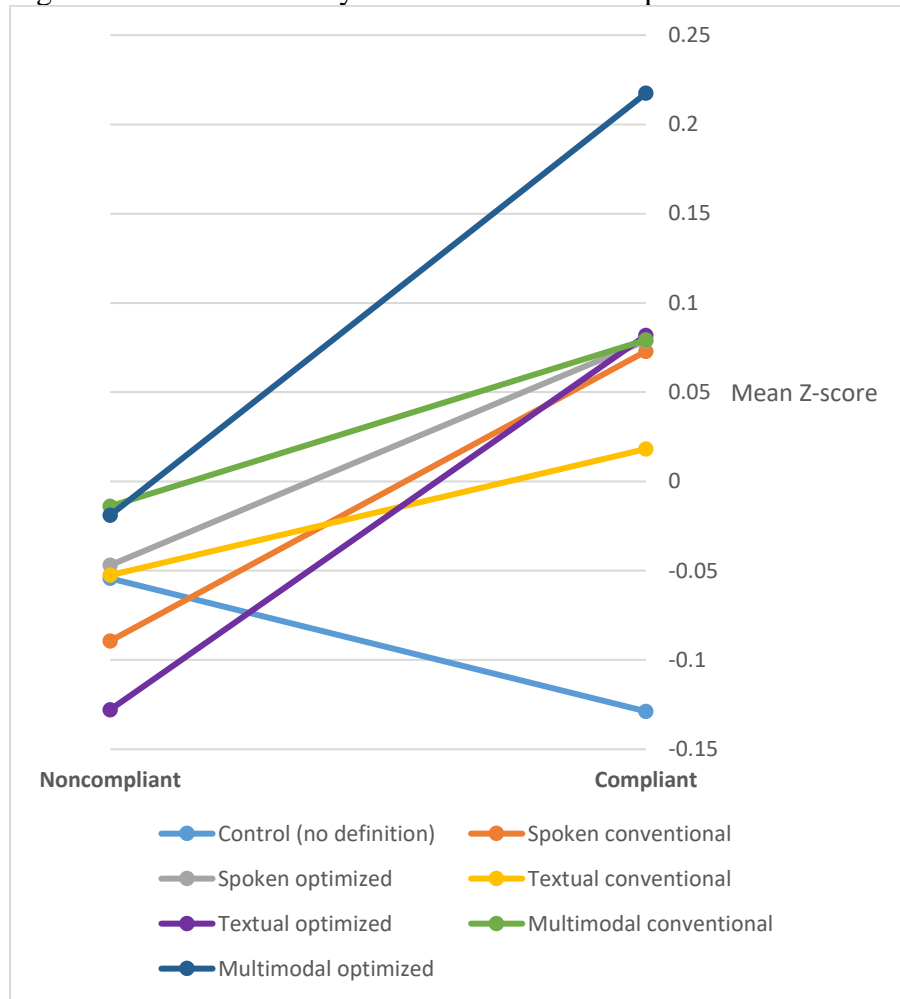
For unimodal definitions, compliant observations were similar for conventional and optimized definitions. For unimodal spoken definitions, the differences between optimized and conventional definitions was not statistically significant ($t(575)=0.12, p=.9077$). The same was true for unimodal textual definitions; the difference between textual optimized and textual conventional definitions was not statistically significant ($t(868)=0.91, p=.3642$).

Finally, whether a given person-item was listened to or read can be added to the models above as in interaction with definition mode to determine whether compliance plays a different role across different definition modes, and to determine whether compliance explains differences (that is, whether mode remains a relevant predictor).

While mode is not significant overall when the model also includes an interaction between compliance and mode ($F(6,1782)=1.61, p=.1395$), some pairwise comparisons still were significant: multimodal optimized definitions yield responses that are significantly more aligned with definitions than spoken conventional ($t(1399)=2.16, p=.0306$) and textual optimized definitions ($t(2440)=2.06, p=.0394$).

The interaction between compliance and definition mode was significant overall ($F(7,3340)=12.39, p<.0001$). This interaction was not significant for textual conventional definitions ($t(6071)=1.34, p=.1790$), possibly because the relatively long estimated reading time may misclassify some observations where definitions were partially, but not fully, read by respondents. For other definition types, compliance was significantly associated with alignment. In particular, compliance led to more alignment under spoken conventional ($t(2695)=4.42, p<.0001$), spoken optimized ($t(2321)=3.18, p=.0015$), textual optimized ($t(6133)=3.56, p=.0004$), multimodal conventional ($t(2171)=2.13, p=.0336$), and multimodal optimized presentations ($t(2709)=6.39, p<.0001$). The average Z-scores for compliant and noncompliant responses by definition mode are shown in Figure 2.

Figure 2- Mean Z-score by definition mode and optimization and compliance status



The magnitude of compliance effects by definition mode can be seen in Figure 2. As noted above, the difference between compliant and noncompliant observations for the control and textual conventional groups was not statistically significant. For all other definition treatments, compliant responses were more aligned with definitions than their noncompliant counterparts. In particular, compliant spoken conventional and spoken optimized definitions were about 0.16 and 0.13 standard deviations higher, respectively, than noncompliant responses. For textual optimized definitions, compliant responses were about 0.21 standard deviations higher than noncompliant responses. For multimodal conventional and optimized definitions, compliant

responses were about 0.09 and 0.24 standard deviations higher, respectively, than their noncompliant counterparts.

4.2.3 Intent to treat

As discussed in 4.1.3 Intent to treat, the effectiveness of definitions can be examined by focusing on the treatments observations effectively received, rather than the treatment to which they were assigned. This section estimates the definition treatment observations received and analyzes those groups, in some cases, allowing respondents to behave as if they were in a different treatment than the one to which they were assigned (within the parameters of the study design). Because compliance is measured for each observation in each treatment, this analysis by “treatment received” is also conducted at the observation level. The average Z-score for treatment received, by definition mode and optimization, is below in Table 20.

Table 20- Mean Z-score by definition mode received and optimization

Definition mode	Mean Z-score
Control (no definition)	-.07063
Spoken conventional	.07089
Spoken optimized	.07892
Textual conventional	.04059
Textual optimized	.09334
Multimodal conventional	.07927
Multimodal optimized	.21742

For responses that complied with the multimodal optimized³ treatment to which they were assigned, answers were more aligned with definitions by about 0.22 standard deviations than the average response, while responses that complied with multimodal conventional definitions were more aligned with definitions by about .08 standard deviations than the average response. Responses that complied with spoken conventional and optimized definitions were more aligned with underlying concepts by about .07 and .08 standard deviations, respectively. Responses that complied with textual spoken and conventional definitions (whether they were assigned to those groups and complied or were assigned to a multimodal group but complied only with instructions for textual definitions) were more aligned with underlying concepts by about .04 and .09 standard deviations, respectively.

³ Observations from multimodal optimized definitions that both read and listened to definitions were considered fully compliant. The average Z-scores for this group are identical whether noncompliant observations are removed from analysis, such as in Table 19, or grouped with another definition type, such as in Table 20. This also holds for observations from multimodal conventional definitions.

Observations compliant with multimodal optimized definitions were significantly more aligned with definitions than each other types of definition. That is, average Z-scores were higher for the multimodal optimized group than when the effective treatment received was multimodal conventional ($t(1566)=2.87, p=.0041$), spoken conventional ($t(1585)=3.55, p=.0004$), spoken optimized ($t(1467)=3.34, p=.0008$), textual conventional ($t(2517)=3.80, p=.0001$), or textual optimized ($t(2471)=3.29, p=.0010$) definitions, as well as the control treatment with no definitions ($t(2471)=8.99, p<.0001$). For treatment actually received, optimization did not lead to statistically significant differences for unimodal definitions. Comparing responses when the effective treatment group was optimized or conventional, there was no difference for either spoken ($t(1412)=0.17, p=.8672$) or textual ($t(3050)=1.09, p=.2773$) definitions.

For observations for which the treatment effectively received was unimodal, the difference between responses that were assigned to a unimodal treatment and complied and responses that were assigned to a multimodal treatment group and did not comply was not statistically significant. That is, compliant unimodal and noncompliant multimodal responses were similar, and if the effective treatment received was unimodal either because respondents complied with their assigned unimodal definitions or because respondents did not comply with their assigned multimodal definitions, there was no difference in responses based on that distinction. This held for both textual conventional definitions ($t(113)=0.58, p=.5617$) and for textual optimized definitions ($t(147)=0.43, p=.6686$).

4.3 Definition inclusivity

Experimental questions and definitions were designed so that using definitions would likely either increase (“inclusive” questions) or decrease (“exclusive” questions) respondents’ numerical answers. For the above analyses, both types of questions were pooled together. However, question type should be examined to determine whether results vary between the 5 inclusive and 7 exclusive definitions. For example, it is possible that multimodal communication is more effective in conveying certain kinds of information, for example, requesting enumeration of additional events (for inclusive definitions) or asking for omission of examples that respondents may have already considered (for exclusive definitions).

The impact of definition inclusivity can be evaluated as a main effect or as an interaction with definition mode. However, because the same questions were asked of all respondents in the same order (beginning with questions with exclusive definitions), definition inclusivity wholly overlaps with question number. Earlier analyses showed that question number explained very little variation in responses, so only the inclusivity or exclusivity of definitions, rather than individual question numbers, are used in these analyses.

Overall, definition exclusivity appears to be related to the degree to which responses were aligned with definitions. For the four definition mode groups (control, unimodal spoken, unimodal textual, multimodal), exclusive definitions led to responses about 0.035 standard deviations higher than inclusive definitions ($t(1100)=2.14, p=.0327$). The same small but statistically significant difference is present when responses are further divided by optimization. However, as noted earlier, compliance varied by survey item and was highest for the first few items in the questionnaire. The first few items of the questionnaire used exclusive, rather than inclusive, definitions, so compliance may explain some differences by question type. The effects of definition inclusivity and question order cannot be disentangled in this design, but warrant further exploration in a setting in which question order can be randomized or rotated.

Comparing responses for only compliant observations, inclusivity of definitions remains a statistically significant predictor of the degree to which responses are consistent with definitions. However, limiting analysis to only compliant observations

reversed the direction of this difference. With compliant observations, responses to questions presented with exclusive definitions were less aligned with definitions by about 0.05 standard deviations when controlling for definition mode ($t(5188)=-2.10$, $p=.0355$) or both definition mode and optimization ($t(5186)=-2.07$, $p=.0387$). This reversal suggests that while whether a definition is inclusive or exclusive may be related to data quality (with inclusive definitions more effective than exclusive), ensuring compliance is more important for collecting data consistent with definitions.

4.4 Respondent characteristics

This section compares the effectiveness of multimodal definitions by respondent characteristics. For each demographic characteristic, alignment with definitions is compared by definition mode (i.e., unimodal spoken, unimodal textual, multimodal) and again by mode and optimization (e.g., multimodal optimized). Some demographic characteristics are expected to have a collinear relationship, for example, age and technology use, so multivariate models are used. These relationships between demographic characteristics and alignment with definitions are considered overall, and then specifically for multimodal and multimodal optimized definitions.

4.4.1 Gender

There was no reason to believe *a priori* that respondent gender would be associated with the degree to which their responses aligned with definitions. However, gender was associated with definition alignment when considering either definition mode only ($F(2,1200)=11.27$, $p<.0001$), or both mode and optimization ($F(2,1200)=11.97$,

$p < .0001$). In particular, responses from men were more aligned with definitions than responses from women by about 0.08 standard deviations. This pattern was slightly weaker when considering multimodal definitions only; for any type of multimodal definition, male responses were more aligned with definitions by about 0.06 standard deviations ($t(355)=1.98, p=.0489$), but the difference of 0.08 standard deviations was not significant at the $p=.05$ level for multimodal optimized definitions ($t(199)=1.96, p=.0511$). Since there is no underlying theory behind these findings, it is unclear whether these differences may be unique to the sample or applicable to a more general population.

4.4.2 Age

Older respondents were shown to be more compliant with definitions, so it is expected that their responses would be more aligned with definitions. However, age was not a significant predictor of definition alignment when considering definition mode ($t(1200)=1.58, p=.1139$) or both mode and optimization ($t(1200)=1.64, p=.1006$). Age remained irrelevant for any type of multimodal definition ($t(355)=0.56, p=.5739$) and for multimodal optimized definitions ($t(201)=1.83, p=.0686$), though the latter comparison may be underpowered.

4.4.3 Race and ethnicity

Like with gender, no particular relationship was expected between the race and ethnicity of respondents and the degree to which responses aligned with definitions. Earlier analysis suggested that race/ethnicity was not related to compliance or, that if a relationship was present, non-White or Hispanic respondents were more likely to

comply with definitions. However, responses from non-White or Hispanic respondents were less aligned with definitions by about 0.9 standard deviations when considering definition mode alone ($t(1200)=4.84, p<.0001$) or both mode and optimization ($t(1200)=4.74, p<.0001$). This surprising pattern persisted for any type of multimodal definition ($t(360)=2.54, p=.0114$) and for multimodal optimized definitions ($t(203)=2.22, p=.0277$).

4.4.4 Education

Higher levels of education were associated with higher levels of compliance, so it is expected that this would also lead to higher levels of alignment with definitions. This pattern emerged when considering definition mode only ($F(3,1200)=5.05, p=.0017$) and both mode and optimization ($F(3,1200)=4.82, p=.0024$). However, this effect did not carry over to multimodal definitions; definition alignment was unrelated to respondents' education for any type of multimodal definition ($F(3,357)=1.50, p=.2129$) and for multimodal optimized definitions ($F(3,200)=2.08, p=.1039$). It is unclear why increased compliance among respondents with more formal education led to more alignment with definitions overall, but not with multimodal definitions.

4.4.5 Technology use

More experience with technology was associated with less compliance with definitions, so it is expected to also be associated with less alignment with definitions. However, respondents with moderate and high levels of technology use provided responses more aligned with definitions by about 0.05 to 0.06 standard deviations than responses from those with the least experience, depending on whether only

definition mode ($F(2,1200)=3.29, p=.0374$) or both mode and optimization ($F(2,1200)=3.16, p=.0423$) are considered. For any type of multimodal definitions, the relationship was not significant ($F(356)=1.41, p=.2446$). However, multimodal optimized definitions appear to drive the overall effect. Multimodal optimized respondents who reported moderate and high levels of technology use provided responses more aligned with definitions by 0.14 ($t(200)=2.46, p=.0147$) and 0.13 ($t(199)=2.13, p=.0342$) standard deviations, respectively.

4.5 Summary

Observations in response to multimodal definitions were more aligned with definitions than observations in response to unimodal definitions, consistent with the primary hypothesis of this research (H1). The effectiveness of multimodal definitions was driven by multimodal optimized definitions, suggesting that only complementary, rather than redundant multimodal content, is effective (and supporting the theory of the redundancy effect, where conveying identical information through multiple channels can reduce comprehension). The increased definition alignment with multimodal, and particularly multimodal optimized definitions, appeared when comparing either all observations or all compliant observations, as compliance with definitions was associated with more alignment with definitions. However, optimization led to increased alignment only for multimodal communication; for unimodal spoken and unimodal textual definitions, there was no difference in definition alignment for conventional or optimized presentations (contradictory to H2).

Responses to questions with exclusive definitions were more aligned with presented definitions than responses to questions with inclusive definitions, however, these results may be driven by the particular question order that we used. When accounting for respondent compliance, responses to questions with inclusive definitions appear to be more aligned with definitions.

It was expected that demographic characteristics associated with higher levels of compliance with definitions would also be associated with responses' alignment with definitions. However, observations from male, and White, non-Hispanic respondents appear to be more aligned with definitions, which is unexplained by compliance. While respondent age was associated with compliance, it was unrelated to alignment with definitions. Surprisingly, although higher levels of education and lower levels of technology use were associated with compliance, this did not necessarily result in more alignment with definitions; for multimodal definitions, there was no relationship between education and alignment with definitions. Higher levels of technology use, rather than lower levels, were associated with more alignment with definitions. While these demographic patterns warrant further research to determine whether they are specific to this study population, analyses strongly indicate that multimodal definitions, and particularly multimodal optimized definitions, can increase response alignment with definitions.

Chapter 5: Respondent burden and satisfaction

This chapter discusses respondents' burden (both objective burden and perceived burden) using different types of definitions, as well as their satisfaction with the survey process. If respondents consider a survey to be too long, too burdensome, or otherwise unsatisfactory, they may choose to break off or provide low effort responses. Respondent acceptance of multimodal communication, particularly compared to expected communication norms (such as a text-only web survey), is crucial before multimodal definitions could be used in a non-experimental setting. If respondents have a strong dislike of multimodal communication or it otherwise produces negative reactions, their perceptions would have to be weighed against the increase in data quality from multimodal definitions.

5.1 Respondent burden

Burden can be evaluated either objectively (time to complete) or subjectively (using self-reported data). Both types of measures are discussed in this section. While researchers may consider a longer survey to be more burdensome than a shorter survey, respondents may not necessarily agree.

5.1.1 Time to complete

The amount of time spent on experimental questions can serve as a proxy for respondent burden, even if respondents do not always fully attend to material shown on a computer screen. Respondents may not describe this time as burdensome, for example, if they find the task engaging, so a longer survey time does not necessarily

indicate that respondents feel burdened. However, analyzing the time spent on survey tasks is important for understanding the scope of a respondent request. In this analysis, we compare the time spent on a questionnaire page before submitting responses for the 12 experimental questions across treatment groups. Table 21 shows the 25th, 50th, and 75th percentiles of this time, as well as the mean time and its standard deviation by definition type.

Table 21- Time spent on 12 definition questions (in seconds)

Definition mode	25 th percentile	Median	75 th percentile	Mean	SD
Control (no definition)	72	93	136	107	51
Spoken conventional	142	288	410	299	205
Spoken optimized	133	198	279	237	175
Textual conventional	90	130	209	169	120
Textual optimized	105	160	195	192	219
Multimodal conventional	140	222	363	264	171
Multimodal optimized	142	192	281	249	235

If respondents attend to all aspects of definitions, then the time spent on the survey is expected to be longest for respondents with spoken components, particularly for spoken conventional and multimodal conventional definitions given the relative length of spoken components (see Table 4), and the time spent on conventional definitions is expected to be longer than their optimized counterparts. This appears to be partially true in Table 21, for example, the mean time to complete textual optimized questions was shorter than multimodal optimized (192 seconds compared to 249 seconds), though this was also longer than textual conventional (169 seconds). These hypotheses are tested formally, and the mean response time is compared across groups using a Tukey adjustment, with results shown below in Table 22. The notation indicates how the treatment listed in rows compares to treatments in columns, and

whether the row treatment was significantly shorter (S), significantly longer (L), or whether the difference was not statistically significant (n.s.).

Table 22- Comparison of questionnaire timing by treatment

	C	SC	SO	TC	TO	MC	MO
Control (C)	-	S	S	n.s.	S	S	S
Spoken conventional (SC)	L	-	L	L	L	n.s.	n.s.
Spoken optimized (SO)	L	S	-	n.s.	n.s.	n.s.	n.s.
Textual conventional (TC)	n.s.	S	n.s.	-	n.s.	S	S
Textual optimized (TO)	L	S	n.s.	n.s.	-	n.s.	n.s.
Multimodal conventional (MC)	L	n.s.	n.s.	L	n.s.	-	n.s.
Multimodal optimized (MO)	L	n.s.	n.s.	L	n.s.	n.s.	-

The control group with no definitions, unsurprisingly, completed their questionnaires in less time than every definition type with the exception of textual conventional definitions (Table 22). When comparing textual definitions and those with a spoken component (either unimodal or multimodal), as predicted, unimodal textual respondents completed their items in less time than in other conditions (besides the control condition). In particular, the time spent on textual conventional definitions was shorter than spoken conventional, multimodal conventional, and multimodal optimized. For textual optimized, this difference was only significant when compared to spoken conventional. The presence of spoken information does appear to increase the time to complete a questionnaire.

When comparing conventional definitions to their optimized counterparts, only one definition type showed a significant difference in time. Spoken conventional respondents used more time than spoken optimized respondents, increasing respondent burden. However, there was no significant difference in the elapsed times

when comparing textual conventional and optimized respondents or when comparing multimodal conventional and optimized respondents.

It is important to note that for each type of definition presentation, the time spent on the 12 experimental questions was fairly short, with the median times ranging from about 1.5 to 4.5 minutes. It is unclear whether these durations are typical when compared to other surveys of MTurk workers or general populations, as little information is available about typical response times. Time per page is often captured through web paradata, and while it is not a key part of this study, additional data about the length of web surveys among various populations would help provide additional context for the findings in this study.

5.1.2 Self-reported burden

Respondents who were presented with unimodal spoken and multimodal definitions were asked to describe how burdensome they found the process of accessing spoken definitions (Not at all burdensome, slightly burdensome, moderately burdensome, very burdensome, extremely burdensome). This question was designed to measure the effort required to play spoken definitions, and so is not applicable to respondents in the control group, who saw no definitions, or respondents who were assigned to view unimodal textual definitions, since textual definitions appeared by default with no additional action needed from respondents. This self-reported burden is shown below in Table 23, both overall, for unimodal and multimodal definitions, and for conventional and optimized definitions.

Table 23- Self-reported burden by definition mode and optimization

	All	Unimodal spoken	Multimodal	Conventional	Optimized
Not at all burdensome	61%	62%	61%	57%	65%
Slightly burdensome	21%	22%	20%	23%	20%
Moderately burdensome	11%	9%	12%	13%	9%
Very burdensome	4%	4%	4%	4%	4%
Extremely burdensome	3%	3%	3%	4%	2%

Overall, respondents did not indicate that playing definitions was burdensome. Most reported that accessing definitions was not at all burdensome (61 percent) or slightly burdensome (21 percent), while few found the process to be very (4 percent) or extremely (3 percent) burdensome. Similar levels of burden were reported regardless of whether respondents were required to play spoken definitions in order to access clarifying instructions (i.e., assigned to unimodal spoken definitions) or, if they could choose to read textual definitions (i.e., assigned to multimodal definitions), with both groups reporting low levels of burden ($\chi^2(4)=1.3306, p=.8562$). Burden was also compared by definition optimization. While the same respondent action was required to *access* conventional and optimized definitions, more time was needed to fully *listen to* conventional definitions, so respondents may perceive the entire process as more burdensome. However, respondent-reported burden to access conventional and optimized definitions was no different ($\chi^2(4) = 7.5893, p=.1078$).

5.2 Satisfaction

All respondents were asked to rate their overall satisfaction with the survey (Overall, how satisfied were you with your experience when responding to this survey? Very dissatisfied, somewhat dissatisfied, neither dissatisfied nor satisfied, somewhat satisfied, very satisfied). These results are shown below in Table 24.

Table 24- Respondent satisfaction by definition mode and optimization

	All	None	Textual	Spoken	Multimodal	Conventional	Optimized
Very dissatisfied	1%	3%	1%	1%	2%	2%	1%
Somewhat dissatisfied	4%	4%	2%	4%	4%	4%	4%
Neither dissatisfied nor satisfied	14%	21%	19%	11%	13%	15%	13%
Somewhat satisfied	33%	40%	39%	31%	29%	34%	31%
Very satisfied	48%	31%	39%	53%	52%	46%	51%

Respondents provided positive feedback about their survey experience. Almost half (48 percent) were very satisfied, and one-third (33 percent) were somewhat satisfied. The remainder were neither dissatisfied nor satisfied (14 percent), somewhat dissatisfied (4 percent), or very dissatisfied (1 percent). This distribution differed by definition mode ($\chi^2(12)=32.2826, p=.0009$), with relatively higher proportions of unimodal spoken and multimodal respondents reporting they were very satisfied when compared to unimodal textual respondents and those who were not shown definitions (53, 52, 39, and 31 percent, respectively). This difference is somewhat surprising, but it is reassuring that the presence of spoken information did not decrease respondent satisfaction, and in fact, respondents who were presented with

either spoken definitions only or multimodal definitions that include a spoken component reported the highest levels of satisfaction.

Regardless of mode, satisfaction was similar when comparing respondents who saw conventional or optimized definitions ($\chi^2(4)=3.0265$, $p=.5534$). Although optimized definitions were designed to be more appealing to respondents, this was not reflected in their assessments of the questionnaire.

5.3 Summary

Self-reported data from respondents indicates that they were satisfied with the survey process, and including a spoken definition in a web survey, either combined with textual definitions or instead of textual definitions, did not lower respondents' satisfaction with the survey process. While respondents who were presented with spoken definitions in a unimodal or multimodal format spent more time completing the survey than respondents who were shown either unimodal textual definitions or no definitions, this additional time does not seem to have negatively impacted respondents' perceptions of the process (or, any negative effects did not outweigh the \$1 incentive respondents received for completing the survey). Spoken definitions add to respondent objective burden (longer duration), but the additional elapsed time and requirement that respondents click to play audio clips did not seem to leave respondents with a negative (subjective) impression of their survey experience. While these results may be related to the sample used, they suggest that multimodal definitions can be implemented in online surveys without overburdening respondents or otherwise causing a negative survey experience.

Chapter 6: Conclusion

6.1 Summary

This dissertation evaluated the feasibility and effectiveness of multimodal definitions in online surveys. Two different explanations were introduced that may explain the effectiveness of multimodal definitions. On the one hand, if spoken and written text are processed at least somewhat independently, then any multimodal communication would improve comprehension when compared to unimodal communication, since more information would be learned and processed by a respondent. On the other hand, if the redundancy effect leads to decreased comprehension, then only complementary multimodal communication, rather than redundant information, would improve response quality over unimodal communication. The study was designed to evaluate those dueling theories.

Compliance with multimodal definitions was not universal, though neither was compliance with unimodal definitions. For unimodal and multimodal definitions, compliance was higher for optimized than for conventional definitions (though definition length and optimization were confounded, so this finding may be due to the relatively shorter length of optimized definitions). For all definition types, compliance was highest for the first survey item than for subsequent questions, but respondents were willing to play spoken definitions in a survey mode that typically includes only text. While compliance could perhaps increase with shorter or more visually appealing definitions (two features that differentiated conventional and optimized presentations), these findings are promising for the efficacy of multimodal

definitions, particularly given the strict compliance criteria for spoken definitions (i.e., audio clips must have been fully played by respondents).

Observations in response to multimodal definitions were more aligned with definitions than observations in response to unimodal definitions, which is consistent with the primary hypothesis of this research, namely, the ability of multimodal communication to increase response alignment with definitions. However, this relationship was driven by multimodal optimized definitions; multimodal conventional definitions did not increase alignment with definitions beyond unimodal definitions. This suggests that the redundancy effect- the need for the information presented in different modes to be at least somewhat non-overlapping and not fully redundant in order to be improve performance- is applicable to survey response quality.

These findings held true when considering all observations, though this effect was particularly pronounced when analyzing only compliant observations. Even with modest compliance rates (about one-third of multimodal optimized observations were estimated to have full compliance with both spoken and textual components), the effectiveness of multimodal definitions was statistically significant. While increasing respondent compliance may lead to greater alignment with definitions, multimodal definitions can improve data quality even when respondents do not fully comply.

Respondents did not find multimodal definitions to be too burdensome or otherwise unsatisfactory. While surveys with definitions that included a spoken component (that is, unimodal spoken and multimodal) took more time complete than surveys that were textual only, respondents did not perceive the longer surveys to be more burdensome, and they were satisfied with the process regardless of what type of definitions they were presented with. If respondents only minimally complied with multimodal definitions, or if they provided negative feedback about their experiences, those drawbacks would have to be carefully weighed against the increased alignment with definitions for responses to multimodal instructions. Instead, these results suggest that respondents do not find multimodal definitions to be burdensome, are willing to comply with instructions to read and listen to them, and will apply these definitions to their survey responses. In an online survey, multimodal definitions can improve data quality without negatively impacting respondents.

6.2 Future research

As noted in discussions of respondent compliance (see 3.1 Compliance with spoken definitions and 3.2 Compliance with textual definitions), compliance was inferred without truly knowing whether respondents attended to definitions. For spoken definitions, compliance may have been underestimated for respondents who partially listened to spoken definitions. For textual definitions, compliance may have been over- or under-estimated if respondent reading speed was miscalculated, and compliance may have been over-estimated for respondents who looked at a different part of their screen. For spoken definitions, a more robust tracking mechanism could assess how much of spoken definitions were played. For textual definitions, a lab

study that tracks respondents' eye movements could more accurately measure whether on-screen text was read.

The sample for this study was drawn from Amazon Mechanical Turk. This sample provided a proof-of-concept that multimodal definitions can improve data quality, but more research is needed to determine the degree to which these findings can be applied to a more general population. In particular, the findings related to technology (higher rates of compliance and less alignment with definitions for observations from respondents with relatively less technology experience) may not hold for a population with more heterogeneous technology experience, or for other specialized populations that are highly different from the typical Mechanical Turk workers.

This study focuses on a fundamentally visual type of survey: a textual web survey, in which spoken definitions were embedded in some experimental conditions. While text is persistent, spoken communication is ephemeral, so improvements in data quality due to adding text to a communication format that is typically spoken (such as telephone surveys) is likely to be greater than the improvements due to adding spoken information to a communication format that is typically textual (such as web surveys). While some spoken surveys do have an added textual component (e.g., show cards), this has typically included response options, rather than questions and definitions. Telephone surveys rarely include a textual component, and this gap is particularly ripe for exploration. Respondents completing a telephone survey are often using an internet-enabled device. A respondent could receive text instructions

from an interviewer, e.g., via a text message, particularly for survey items for which instructions are nuanced or potentially counter-intuitive. This type of technology has been integrated into customer service interactions, for example, customers calling the U.S. Postal Service can consent to receive information about their call by text and even submit responses by text message, as a substitute or supplement to interacting with an IVR system⁴. While the effectiveness of multimodal communication may differ across these applications, particularly given differences in communication norms and respondent expectations, these uses warrant further exploration of multimodal definitions given respondent expectations for communication, technological advances, and the potential to improve data quality.

⁴ More information at <https://faq.usps.com/s/article/How-do-I-Navigate-the-Interactive-Voice-Response-IVR-System>

Appendix A- Questionnaire

This survey should be completed in Chrome or Firefox.

Please read/listen/both read and listen to all instructions and definitions carefully before answering each question. This will help ensure that you provide the best information you can.

1. In the past 7 days, how many hours of television did you watch?
2. In the past 7 days, for how many hours did you listen to the radio?
3. In the past 7 days, how many e-mails did you send?
4. In the past 7 days, for how many hours did you use e-mail?
5. Excluding e-mail use, in the past 7 days, for how many hours did you use the internet?
6. In the past 7 days, how many text messages did you receive?
7. In the past 7 days, how many hours did you work in total?
8. In the past 7 days, how many miles did you travel by vehicle?
9. In the past year, how many plane trips did you take?
10. In the past 30 days, how many times have you had food or drinks at a restaurant?
11. How many pairs of shoes do you own?
12. How many short-sleeved t-shirts do you own?
13. How many hours of rest do you get on a typical weekday?
14. In the past 7 days, how many hours did you exercise?
15. In the past 7 days, how many caffeinated drinks did you have?
16. List all food you've eaten in meals or snacks in the past 3 days that included any form of poultry.
17. List all food you've eaten in meals or snacks in the past 3 days that included any form of dairy products.

18. List all food you've eaten in meals or snacks in the past 3 days that included any form of vegetables.
19. For unimodal textual and multimodal respondents: For how many questions did you read the provided definitions? None, a few, most, all
20. For unimodal spoken and multimodal respondents: For how many questions did you listen to the provided definitions? None, a few, most, all
21. For unimodal spoken and multimodal respondents: For how many questions did you speed up the playback of the provided definitions? None, a few, most, all
22. For unimodal spoken and multimodal respondents: For how many questions did you mute your computer's sound? None, a few, most, all
23. For multimodal respondents: For how many questions did you both read and listen to the provided definitions? None, a few, most, all
24. For multimodal respondents who both read and listened: Which of the following best describes the order in which you read and listened to definitions? I read before I listened; I listened before I read; I read and listened simultaneously
25. For respondents who accessed any definitions: How useful were definitions in answering survey items? Not at all useful, slightly useful, moderately useful, very useful, extremely useful
26. For unimodal textual and multimodal respondents: Why did you choose to read (or not read) definitions?
27. For unimodal spoken and multimodal respondents: Why did you choose to listen to (or not listen to) definitions?
28. For spoken and multimodal respondents: How burdensome was it to access spoken definitions? Not at all burdensome, slightly burdensome, very burdensome
29. For on-demand respondents: How satisfied were you with the steps needed to access definitions? Very dissatisfied, somewhat dissatisfied, neither dissatisfied nor satisfied, somewhat satisfied, very satisfied
30. Overall, how satisfied were you with your experience when responding to this survey? Very dissatisfied, somewhat dissatisfied, neither dissatisfied nor satisfied, somewhat satisfied, very satisfied
31. Which of the following best describes you?
 - Male

- Female
- Another gender

32. Which of the following best describe your race/ethnicity? Select all that apply.

- ☐ American Indian/Native American
- ☐ Asian
- ☐ Black/African American
- ☐ Hispanic/Latino
- ☐ White/Caucasian
- ☐ Pacific Islander

33. In what year were you born?

34. What is your highest level of education?

- Some high school
- High school diploma/GED
- Some college
- Associate's degree
- Bachelor's degree
- Master's degree or higher

And now a few questions about your use of technology:

35. How often do you send and receive text messages on a mobile phone?

- Never
- Several times a month
- Once a week
- Several times a week
- Once a day

- Several times a day or more
- 36. How often do you use apps (for any purpose) on a mobile phone?
 - Never
 - Several times a month
 - Once a week
 - Several times a week
 - Once a day
 - Several times a day or more
- 37. How often do you watch TV shows, movies, etc. on a computer?
 - Never
 - Several times a month
 - Once a week
 - Several times a week
 - Once a day
 - Several times a day or more
- 38. How often do you search the Internet for information on any device?
 - Never
 - Several times a month
 - Once a week
 - Several times a week
 - Once a day
 - Several times a day or more

Depending on the experimental condition to which respondents are assigned, they may be exposed to (or have the option of clicking to be exposed to) the following textual and spoken definitions throughout the survey:

QUESTION	CONVENTIONAL DEFINITION	OPTIMIZED (TEXTUAL)	OPTIMIZED (SPOKEN)	INCLUSIVE/EXCLUSIVE
IN THE PAST 7 DAYS, HOW MANY HOURS OF TELEVISION DID YOU WATCH?	Watching television includes any programs other than films. This may include sitcoms, dramas, news, sports, and reality shows. Television is watched on a television set at the time it is broadcast and does not include programming recorded with a DVR, viewed on-demand, or streamed. Include content viewed on a television set only and exclude any content viewed on a computer or mobile device.	<ul style="list-style-type: none"> • Content is broadcast. Exclude DVRed, on-demand, and streamed shows. • TV set. Exclude shows watched on a computer or mobile device. • TV shows. Exclude films, even if watched while they air. 	By television, we mean content watched on a TV set at the time it is broadcast. Exclude streamed, on –demand, and DVRed shows and anything watched on a computer or mobile device. Exclude films.	Exclusive definition

QUESTION	CONVENTIONAL DEFINITION	OPTIMIZED (TEXTUAL)	OPTIMIZED (SPOKEN)	INCLUSIVE/EXCLUSIVE
IN THE PAST 7 DAYS, FOR HOW MANY HOURS DID YOU LISTEN TO THE RADIO?	Listening to the radio includes listening to programming transmitted and received through an antenna. Available stations and reception are restricted by signal strength and listener location. Programs are listened to live, that is, as they air, rather than played later by the listener, such as with podcasts and other downloadable content. Programming can include talk-based content, such as news or sports, but does not include music even if accessed by antenna.	<ul style="list-style-type: none"> • Antenna. Only count local stations through over-the-air access, not satellite or internet. • Live Content. Exclude podcasts or other content played on-demand. • Talk. Programming includes news, sports, and talk shows. Exclude music. 	By radio, we mean local programming listened to live, over-the-air, but not podcasts, satellite radio, or internet radio. Include news, sports, and talk shows, not music.	Exclusive definition

QUESTION	CONVENTIONAL DEFINITION	OPTIMIZED (TEXTUAL)	OPTIMIZED (SPOKEN)	INCLUSIVE/EXCLUSIVE
IN THE PAST 7 DAYS, HOW MANY E-MAILS DID YOU SEND?	E-mail is a message distributed by electronic means from one computer user to one or more recipients via a network. E-mails include a header, with fields such as a subject line, sender, recipient, and date, as well as information contained in the body of the message. The body may include text, as well as embedded or attached multimedia content.	<ul style="list-style-type: none"> • Distribution. Electronic communication to recipient(s). • Components. Header (including subject line, sender, recipient, date) and body (text, multimedia content). 	By e-mail, we mean electronic messages sent to one or more recipients. An e-mail has a header with a subject line, sender, recipient, and date. An e-mail also has a body, with text or multimedia content.	Neutral definition

QUESTION	CONVENTIONAL DEFINITION	OPTIMIZED (TEXTUAL)	OPTIMIZED (SPOKEN)	INCLUSIVE/EXCLUSIVE
IN THE PAST 7 DAYS, FOR HOW MANY HOURS DID YOU USE E-MAIL?	E-mail use includes composing, sending, and reading messages, as well as managing an inbox. Count time spent using an online mailbox, desktop mailbox, or mobile application, and do not count time spent reading attachments or linked content in a browser. Only count e-mail use when connected to the internet through a wired or wireless (Wi-Fi) connection. Exclude email use involving a cellular connection such as 3G or 4G. Exclude offline use.	<p>Exclude</p> <ul style="list-style-type: none"> E-mail using a cellular network such as 3G or 4G. Reading attachments or linked content. <p>Include</p> <ul style="list-style-type: none"> Composing, sending, reading, and sorting messages. Use of a Wi-fi or wired connection. 	By e-mail use, we mean writing, reading, sending and sorting messages. Only count time using an application, not time spent reading attachments or linked content. Only count access through a wired or Wi-Fi connection, so exclude cellular networks such as 3G and 4G.	Exclusive definition

QUESTION	CONVENTIONAL DEFINITION	OPTIMIZED (TEXTUAL)	OPTIMIZED (SPOKEN)	INCLUSIVE/EXCLUSIVE
EXCLUDING E-MAIL USE, IN THE PAST 7 DAYS, FOR HOW MANY HOURS DID YOU USE THE INTERNET?	People may use the Internet to carry out personal or professional tasks and activities. Exclude internet use involving a cellular connection such as 3G or 4G. Include active tasks such as reading news articles, posting in online forums, and playing online games. Exclude passive tasks that do not involve direct attention or engagement such as streaming videos or music.	<ul style="list-style-type: none"> • Connection. Count Wi-fi and wired connections only. Exclude cellular networks such as 3G and 4G. • Active use. Count tasks such reading articles, posting in forums, and playing online games. Do not count passive activities such as streaming videos or music. 	By Internet, we mean access through a wired or Wi-Fi connection, so exclude cellular networks such as 3G and 4G. Only count time on tasks such as reading or posting content or playing games, and do not count passive activities such as streaming videos or music.	Exclusive definition

QUESTION	CONVENTIONAL DEFINITION	OPTIMIZED (TEXTUAL)	OPTIMIZED (SPOKEN)	INCLUSIVE/EXCLUSIVE
IN THE PAST 7 DAYS, HOW MANY TEXT MESSAGES DID YOU RECEIVE?	Text messages are short electronic messages that typically consist of alphabetic and numeric characters. Text messages may also contain digital images such as emojis, photos, and other icons. These are usually sent between mobile phones, but may also be used for communication between users on tablets or other devices. These may be sent or received over either a cellular network or using an internet connection. Text messages can be used to communicate with a single individual, a group of individuals, or an automated system.	<ul style="list-style-type: none"> • Length. Short electronic messages. • Content. Include text and/or images, such as emojis. • Device. Sent from or received on a mobile phone, tablet, or other device. • Network. Sent or received via cellular network or through the internet. • Users. Sent from a single sender or from a group chat. 	By text messages, we mean short electronic communication. These may contain text only, but can also include images such as emojis. Text messages are often sent through mobile phones, but can be sent between users on other devices through a cellular network or using an internet connection. Count messages sent just to you or to a group.	Neutral definition

QUESTION	CONVENTIONAL DEFINITION	OPTIMIZED (TEXTUAL)	OPTIMIZED (SPOKEN)	INCLUSIVE/EXCLUSIVE
IN THE PAST 7 DAYS, HOW MANY HOURS DID YOU WORK IN TOTAL?	Work is paid employment performed for an employer or, if self-employed, for oneself. Count paid internships or apprenticeships. Count time directly spent on work activities, such as time at an office or work site, as well as commuting to and from an office.	Include <ul style="list-style-type: none"> • Paid work or self-employment. • Work as an employee or paid intern. • Time at work and commuting to and from work. 	By work, we mean a paid job or internship, or self-employment. In addition to time at a job site, work includes commuting time.	Inclusive definition
IN THE PAST 7 DAYS, HOW MANY MILES DID YOU TRAVEL BY VEHICLE?	Vehicles have two or more wheels, are used for ground transportation and can include cars, trucks, taxis, buses, trains, subways, trams, motorcycles, and bicycles. All miles spent in a vehicle, regardless of seat location, should be considered. Miles as both a driver and passenger should be included.	<ul style="list-style-type: none"> • Vehicle. Count any ground travel by vehicle, including cars, trucks, taxis, buses, motorcycles, trains, subways, and bicycles. • Role. Count miles as both driver and passenger. 	By travel, we mean miles as a driver or passenger in a vehicle such as a car, truck, taxi, bus, train, subway, tram, motorcycle, or bicycle.	Inclusive definition

QUESTION	CONVENTIONAL DEFINITION	OPTIMIZED (TEXTUAL)	OPTIMIZED (SPOKEN)	INCLUSIVE/EXCLUSIVE
IN THE PAST YEAR, HOW MANY PLANE TRIPS DID YOU TAKE?	A plane trip begins at liftoff and ends at touchdown. If multiple legs (liftoffs and touchdowns) are involved, such as with non-direct or multi-city flights, each is counted separately. Similarly, for roundtrip flights, outbound and return flights are each counted separately, and all legs are counted separately.	<ul style="list-style-type: none"> • Count each leg of a trip separately. • Count roundtrip flights separately. 	Count each component of a trip separately. For example, layovers and roundtrip flights should be counted as multiple plane trips.	Inclusive definition

QUESTION	CONVENTIONAL DEFINITION	OPTIMIZED (TEXTUAL)	OPTIMIZED (SPOKEN)	INCLUSIVE/EXCLUSIVE
IN THE PAST 30 DAYS, HOW MANY TIMES HAVE YOU HAD FOOD OR DRINKS AT A RESTAURANT?	Restaurants are dining establishments at which food and/or beverages are served. Include sit-down establishments, restaurants with and without table service, fast food restaurants, coffee shops and cafes, bars and pubs, food trucks, and street vendors. Food may be eaten at the restaurant or elsewhere, if ordered for take-out, to-go, or delivery.	<ul style="list-style-type: none"> • Type. Count sit-down restaurants, fast food, coffee shops, bars, food trucks and street vendors. • Location. Count dine-in, take-out, to-go orders, and delivery. 	We mean sit-down restaurants, fast food, coffee shops, bars, food trucks and street vendors. We mean dine-in, take-out, to-go orders, and delivery.	Inclusive definition

QUESTION	CONVENTIONAL DEFINITION	OPTIMIZED (TEXTUAL)	OPTIMIZED (SPOKEN)	INCLUSIVE/EXCLUSIVE
HOW MANY PAIRS OF SHOES DO YOU OWN?	Shoes are footwear worn primarily outdoors and secured to a foot with some type of fastener, such as laces, zipper, Velcro, clasps, or buckles. For this question, footwear designed primarily for indoor use such as slippers does not qualify. For this question, non-fastening shoes such as flip flops, slides, clogs, pumps, and other unsecured footwear do not qualify.	<p>Exclude shoes</p> <ul style="list-style-type: none"> Worn indoors, including slippers. Unsecured, such as flip flops, slides, clogs, pumps, etc. <p>Include shoes</p> <ul style="list-style-type: none"> Worn outside Secured with laces, zippers, Velcro, clasps, buckles, etc. 	By shoes, we mean footwear worn primarily outside that can be secured with fasteners such as laces, zippers, Velcro, clasps, or buckles. Do not count unsecured footwear such as flip flops, slides, clogs, pumps, and other unsecured footwear.	Exclusive definition

QUESTION	CONVENTIONAL DEFINITION	OPTIMIZED (TEXTUAL)	OPTIMIZED (SPOKEN)	INCLUSIVE/EXCLUSIVE
HOW MANY SHORT-SLEEVED T-SHIRTS DO YOU OWN?	A short-sleeved t-shirt is a fabric shirt with short sleeves (creating a “T” shape) and neither a collar nor buttons. T-shirts may have round necklines, also known as a crew neck, or v-shaped necklines, also known as a v-neck. T-shirts are usually made out of cotton or some other light and inexpensive fabric.	<ul style="list-style-type: none"> • Design. Short sleeves and no collar or buttons. May feature round or v-shaped neckline. • Material. Usually made of cotton or other light, inexpensive fabric. 	By short-sleeved t-shirt, we mean short-sleeved shirts without a collar or buttons. T-shirts have round or v-shaped necklines and are usually made out of cotton or similar light and inexpensive fabrics.	Neutral definition

**HOW MANY HOURS
OF REST DO YOU
GET ON A TYPICAL
WEEKDAY?**

Include time spent in a state of sleep or time that has the potential to become sleep. This includes overnight sleep and daytime naps, as well as time when sleep is not necessarily intended, such as during class or a meeting, while reading a book, or while watching television.

- **Time of day.** Count evening and daytime rest.
- **Sleep state.** Count time spent asleep or when sleep is possible, such as sitting while reading a book or watching television.

By rest, we mean time when you are asleep or could fall asleep, such as sitting while reading a book or watching TV.

Inclusive definition

QUESTION	CONVENTIONAL DEFINITION	OPTIMIZED (TEXTUAL)	OPTIMIZED (SPOKEN)	INCLUSIVE/EXCLUSIVE
IN THE PAST 7 DAYS, HOW MANY HOURS DID YOU EXERCISE?	<p>Exercise is physical activity that results in an elevated heart rate. This can included vigorous activities such as running or biking and less vigorous activities such as walking, climbing up or down stairs, and yoga. Exercise can be performed alone, such as swimming or biking, or with a group or team, such as basketball or tennis. Include all physical activities, regardless of how long they lasted.</p>	<ul style="list-style-type: none"> • Activities. Count all activities that result in an elevated heart rate. • Duration. Count all physical activities, regardless of how long they lasted. 	By exercise, we mean activities that result in an elevated heart rate, regardless of the duration of each activity.	Inclusive definition

QUESTION	CONVENTIONAL DEFINITION	OPTIMIZED (TEXTUAL)	OPTIMIZED (SPOKEN)	INCLUSIVE/EXCLUSIVE
IN THE PAST 7 DAYS, HOW MANY CAFFEINATED DRINKS DID YOU HAVE?	Caffeine is a stimulant often found in cacao plants and a variety of beverages. Common caffeinated beverages include coffee, tea, and sodas. While caffeinated beverages may be consumed in any amount or container size, for this question, 8 fluid ounces of a caffeinated beverage is one caffeinated drink.	<ul style="list-style-type: none"> Count every 8 ounces as one drink. Count coffee, tea, soda, and other caffeinated beverages. 	By caffeinated drinks, we mean 8 ounces of caffeinated beverages such as coffee, tea, and soda.	Inclusive definition
LIST ALL FOOD YOU'VE EATEN IN MEALS OR SNACKS IN THE PAST 3 DAYS THAT INCLUDED ANY FORM OF POULTRY.	Although often low in fat and cholesterol, when poultry is fried or eaten with its skin, it can be high in fat and cholesterol. Include chickens, turkeys, ducks, geese, and other game birds eaten in any form. Include eggs that come from any of these animals, eaten in any form.	<ul style="list-style-type: none"> Birds. List chickens, turkeys, and other game birds. Eggs. List any meals or dishes made with eggs. 	We mean meals or snacks that include birds, such as chicken or turkey, eggs, and dishes made with eggs.	Inclusive definition

QUESTION	CONVENTIONAL DEFINITION	OPTIMIZED (TEXTUAL)	OPTIMIZED (SPOKEN)	INCLUSIVE/EXCLUSIVE
LIST ALL FOOD YOU'VE EATEN IN MEALS OR SNACKS IN THE PAST 3 DAYS THAT INCLUDED ANY FORM OF DAIRY PRODUCTS.	<i>Dairy products</i> are a type of food produced from or containing the milk of mammals, primarily cattle, water buffaloes, goats, sheep, and camels. Dairy products can be eaten on their own such as milk, cream, ice cream, cheese, and yogurt, or they can be cooked. Dairy products can be combined and eaten with foods from other food groups, for example, creamy soups like cream of mushroom, cheesy foods like pizza and lasagna, and baked goods that were made with butter.	<ul style="list-style-type: none"> • Milk, ice cream, cheese, yogurt, etc. • Creamy soups and other foods made with milk • Pizza, lasagna, and other foods made with cheese • Baked goods and other foods made with butter 	We mean dairy products eaten on their own, such as milk, ice cream, cheese, and yogurt, and foods for which dairy is an ingredient, such as creamy soups, cheese pizza or lasagna, and baked goods.	Inclusive definition

QUESTION	CONVENTIONAL DEFINITION	OPTIMIZED (TEXTUAL)	OPTIMIZED (SPOKEN)	INCLUSIVE/EXCLUSIVE
LIST ALL FOOD YOU'VE EATEN IN MEALS OR SNACKS IN THE PAST 3 DAYS THAT INCLUDED ANY FORM OF VEGETABLES.	Vegetables can be eaten raw or cooked in a variety of ways, such as steamed, roasted, pureed, or fried. Vegetables include the edible stems, leaves, and roots of a plant. Potatoes, including fries, mashed potatoes, and potato chips, are vegetables.	<ul style="list-style-type: none"> • Eaten raw, steamed, roasted, pureed, fried, etc. • List edible stems, leaves and roots • List potato-based dishes, including fries, mashed potatoes, and potato chips 	We mean vegetables eaten raw or cooked in any form, such as steamed, roasted, pureed, or fried. Count potatoes, including fries, mashed potatoes, and potato chips, as well as plant stems, leaves, and roots.	Inclusive definition

Appendix B- Mean Z-score by question and definition mode and optimization

Question	Control	Spoken	Textual	Multimodal	Definition type
1	-0.4691444	0.05923572	0.02478921	0.10589361	Exclusive
2	-0.1016157	0.87210996	0.01665041	0.11718351	Exclusive
3	-0.0128994	-0.002275	0.01402434	-0.079922	Neutral
4	-0.1758259	0.04247729	0.09632973	0.00974737	Exclusive
5	-0.4916486	0.0534586	-0.0979948	0.13222499	Exclusive
6	-0.0422902	-0.0420625	-0.0404427	-0.0118868	Neutral
7	0.26995953	-0.0528783	0.03162846	-0.0424111	Inclusive
8	-0.0607396	-0.0133699	-0.1550516	0.04152386	Inclusive
9	-0.1704922	-0.0606565	-0.0395292	0.05730905	Inclusive
10	-0.0331874	-0.0367362	-0.080249	0.0339123	Inclusive
11	-0.034943	0.0281058	0.13913454	-0.0330795	Exclusive
12	-0.0522887	-0.0735266	-0.0402987	0.03148825	Neutral
13	-0.1755325	-0.0352435	0.03080159	0.03344346	Inclusive
14	-0.0085971	-0.0592645	-0.0343985	-0.0165547	Inclusive
15	-0.064935	-0.0196517	-0.0948726	0.05517593	Inclusive

Question	Control	Spoken conventional	Spoken optimized	Textual conventional	Textual optimized	Multimodal conventional	Multimodal optimized	Definition type
1	-0.4691444	0.01470386	0.11560517	0.00374564	0.05135672	0.00262581	0.18474341	Exclusive
2	-0.1016157	-0.0611636	0.05365152	-0.0562541	0.1086924	0.06439637	0.15832643	Exclusive
3	-0.0128994	-0.0019307	-0.0027	-0.059887	0.10733737	-0.0102643	-0.1337637	Neutral
4	-0.1758259	-0.0068834	0.10250234	0.08384809	0.11212927	-0.1108413	0.10340847	Exclusive
5	-0.4916486	-0.0116917	0.13376966	-0.0751462	-0.126917	0.10837781	0.15078584	Exclusive
6	-0.0422902	-0.0624487	-0.0170202	-0.0160262	-0.0713495	-0.0027151	-0.0189761	Neutral
7	0.26995953	0.06064912	-0.1924959	0.00930764	0.05988265	-0.0350047	-0.0481556	Inclusive
8	-0.0607396	-0.0942135	0.08655471	-0.2044613	-0.0926718	0.12699564	-0.0245413	Inclusive
9	-0.1704922	-0.0669188	-0.0529639	-0.0001545	-0.0892397	0.09618844	0.02696415	Inclusive
10	-0.0331874	-0.0543418	-0.0150008	-0.1486784	0.00614311	-0.0536888	0.10162328	Inclusive
11	-0.034943	0.06172167	-0.0133953	0.16465211	0.10723758	0.02747264	-0.0798831	Exclusive
12	-0.0522887	-0.0769718	-0.0692733	-0.0617328	-0.013238	0.04198964	0.02337124	Neutral
13	-0.1755325	-0.029337	-0.0420339	-0.0456212	0.12148999	-0.1660002	0.19185872	Inclusive
14	-0.0085971	-0.0529956	-0.0670039	-0.0320625	-0.0373477	0.05496368	-0.072103	Inclusive
15	-0.064935	-0.0593908	0.02940884	-0.0774228	-0.116903	0.01900984	0.08313039	Inclusive

Appendix C- Mean response by question and definition mode and optimization

Question	Control	Spoken	Textual	Multimodal	Definition type
1	18.1682692	10.9315642	11.2651934	10.5865922	Exclusive
2	5.73317308	4.39513889	4.19392265	3.22451791	Exclusive
3	28.8173077	32.5662983	30.5414365	23.4686649	Neutral
4	8.60679612	6.04387187	5.39860335	6.48306011	Exclusive
5	42.1666667	30.5098592	33.7486034	28.8254848	Exclusive
6	81.2403846	105.193906	82.1955307	101.324251	Neutral
7	39.3242718	33.5348189	35.0502793	33.7225275	Inclusive
8	122.076923	137.977778	105.220994	149.419619	Inclusive
9	1.05769231	1.59833795	1.5801105	1.997260	Inclusive
10	7.84615385	7.57734807	6.97790055	8.17438692	Inclusive
11	10.8942308	9.89226519	8.57777778	10.5395095	Exclusive
12	20.4230769	19.9558011	22.7955801	22.8964578	Neutral
13	7.1	7.49525316	7.64634146	7.65843949	Inclusive
14	4.84660194	4.59751381	5.05248619	4.9931694	Inclusive
15	11.4230769	11.8149171	11.3922652	12.7138965	Inclusive

NOTE: Units vary by question. See Appendix A- Questionnaire for question wording.

Question	Control	Spoken conventional	Spoken optimized	Textual conventional	Textual optimized	Multimodal conventional	Multimodal optimized	Definition type
1	18.1682692	11.6675	10	11.5445545	10.9125	12.0903226	9.43842365	Exclusive
2	5.73317308	4.83080808	3.86265432	4.92178218	3.275	3.58176101	2.94607843	Exclusive
3	28.8173077	33.475	31.4444444	26.9207921	35.1125	27.8	20.1207729	Neutral
4	8.60679612	6.51142132	5.47530864	5.522	5.24240506	7.88125	5.39708738	Exclusive
5	42.1666667	31.9030612	28.7924528	33.26	34.3670886	29.335443	28.4285714	Exclusive
6	81.2403846	104.914573	105.537037	90.07	72.2278481	116.79375	89.3671498	Neutral
7	39.3242718	35.5707071	31.0310559	34.65	35.556962	33.8553459	33.6195122	Inclusive
8	122.076923	115.673367	165.546584	93.2277228	120.3625	175.39375	129.342995	Inclusive
9	1.05769231	1.50251256	1.71604938	1.68316832	1.45	2.18125	1.85365854	Inclusive
10	7.84615385	7.275	7.95061728	6.36633663	7.75	7.225	8.90821256	Inclusive
11	10.8942308	9.355	10.5555556	8.18	9.075	9.86875	11.057971	Exclusive
12	20.4230769	19.88	20.0493827	23.5940594	21.7875	22.75	23.0096618	Neutral
13	7.1	7.49112426	7.5	7.46067416	7.86666667	7.17985612	8.03857143	Inclusive
14	4.84660194	4.7875	4.36296296	4.6039604	5.61875	5.3625	4.70631068	Inclusive
15	11.4230769	11.38	12.3518519	11.1683168	11.675	12.23125	13.0869565	Inclusive

NOTE: Units vary by question. See Appendix A- Questionnaire for question wording.

Bibliography

- Anderson, Monica. 2019. "Mobile Technology and Home Broadband 2019." *Pew Research Center: Internet, Science & Tech* (blog). June 13, 2019. <https://www.pewresearch.org/internet/2019/06/13/mobile-technology-and-home-broadband-2019/>.
- Baddeley, Alan. 1992. "Working Memory." *Science* 255 (5044): 556–559.
- Carver, Ronald P. 1992. "Reading Rate: Theory, Research, and Practical Implications." *Journal of Reading* 36 (2): 84–95.
- Clark, Herbert H. 1996. "Community, Commonalities, and Communication." In *Rethinking Linguistic Relativity*, edited by John J. Gumperz and Stephen C. Levinson. Cambridge University Press.
- Conrad, Frederick G., Mick P. Couper, Roger Tourangeau, and Andrey Peytchev. 2006. "Use and Non-Use of Clarification Features in Web Surveys." *Journal of Official Statistics* 22 (2): 245.
- Conrad, Frederick G., and Michael F. Schober. 2000. "Clarifying Question Meaning in a Household Telephone Survey." *Public Opinion Quarterly* 64 (1): 1–28. <https://doi.org/10.1086/316757>.
- Conrad, Frederick G., Michael F. Schober, and Tania Coiner. 2007. "Bringing Features of Human Dialogue to Web Surveys." *Applied Cognitive Psychology* 21 (2): 165–87. <https://doi.org/10.1002/acp.1335>.
- Couper, Mick P. 2011. "The Future of Modes of Data Collection." *Public Opinion Quarterly* 75 (5): 889–908. <https://doi.org/10.1093/poq/nfr046>.
- Couper, Mick P., R. Tourangeau, and T. Marvin. 2009. "Taking the Audio Out of Audio-CASI." *Public Opinion Quarterly* 73 (2): 281–303. <https://doi.org/10.1093/poq/nfp025>.
- Dumas, Bruno, Denis Lalanne, and Sharon Oviatt. 2009. "Multimodal Interfaces: A Survey of Principles, Models and Frameworks." In *Human Machine Interaction*, edited by Denis Lalanne and Jürg Kohlas, 3–26. Lecture Notes in Computer Science 5440. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-00437-7_1.
- Galesic, Mirta, Roger Tourangeau, Mick P. Couper, and Frederick G. Conrad. 2008. "Eye-Tracking Data." *Public Opinion Quarterly* 72 (5): 892–913. <https://doi.org/10.1093/poq/nfn059>.
- Jäckle, Annette, Caroline Roberts, and Peter Lynn. 2006. "Telephone versus Face-to-Face Interviewing: Mode Effects on Data Quality and Likely Causes. Report on Phase II of the ESS-Gallup Mixed Mode Methodology Project." ISER Working Paper 2006-41. Colchester: University of Essex.
- Johnston, Michael. 2008. "Automating the Survey Interview with Dynamic Multimodal Interfaces." In *Envisioning the Survey Interview of the Future*, edited by F.G. Conrad and M.F. Schober. Hoboken, NJ: Wiley.
- Kalyuga, Slava, Paul Chandler, and John Sweller. 2004. "When Redundant On-Screen Text in Multimedia Technical Instruction Can Interfere With Learning." *Human Factors: The Journal of the Human Factors and*

- Ergonomics Society* 46 (3): 567–81.
<https://doi.org/10.1518/hfes.46.3.567.50405>.
- Kunz, Tanja, and Marek Fuchs. 2012. “Positioning of Clarification Features in Web Surveys: Evidence from Eye Tracking Data.” In *JSM Proceedings of the Survey Research Methods Section*. Alexandria, VA.
- Leahy, Wayne, and John Sweller. 2011. “Cognitive Load Theory, Modality of Presentation and the Transient Information Effect.” *Applied Cognitive Psychology* 25 (6): 943–51. <https://doi.org/10.1002/acp.1787>.
- Liu, Rong, Bhavika N. Patel, and MiYoung Kwon. 2017. “Age-Related Changes in Crowding and Reading Speed.” *Scientific Reports* 7 (August).
<https://doi.org/10.1038/s41598-017-08652-0>.
- Mayer, Richard E., Julie Heiser, and Steve Lonn. 2001. “Cognitive Constraints on Multimedia Learning: When Presenting More Material Results in Less Understanding.” *Journal of Educational Psychology* 93 (1): 187–98.
<https://doi.org/10.1037/0022-0663.93.1.187>.
- Mayer, Richard E., and Cheryl I. Johnson. 2008. “Revising the Redundancy Principle in Multimedia Learning.” *Journal of Educational Psychology* 100 (2): 380–86. <https://doi.org/10.1037/0022-0663.100.2.380>.
- Metzler, Anke, Tanja Kunz, and Marek Fuchs. 2015. “The Use and Positioning of Clarification Features in Web Surveys.” *Psihologija* 48 (4): 379–408.
- Miller, Peter V. 1984. “Alternative Question Forms for Attitude Scale Questions in Telephone Interviews.” *Public Opinion Quarterly* 48 (4): 766–78.
<https://doi.org/10.1086/268882>.
- Moreno, Roxana, and Richard E. Mayer. 2002. “Verbal Redundancy in Multimedia Learning: When Reading Helps Listening.” *Journal of Educational Psychology* 94 (1): 156–63. <https://doi.org/10.1037/0022-0663.94.1.156>.
- Mousavi, Seyed Yaghoub, Renae Low, and John Sweller. 1995. “Reducing Cognitive Load by Mixing Auditory and Visual Presentation Modes.” *Journal of Educational Psychology* 87 (2): 319–34. <https://doi.org/10.1037/0022-0663.87.2.319>.
- Peytchev, Andy, Frederick G. Conrad, Mick P. Couper, and Roger Tourangeau. 2010. “Increasing Respondents’ Use of Definitions in Web Surveys.” *Journal of Official Statistics* 26 (4): 633–50.
- Redline, Cleo. 2013. “Clarifying Categorical Concepts in a Web Survey.” *Public Opinion Quarterly* 77 (S1): 89–105. <https://doi.org/10.1093/poq/nfs067>.
- Redline, Cleo, Andrew Zukerberg, Chelsea Owens, and Amy Ho. 2016. “Instructions in Self-Administered Survey Questions: Do They Improve Data Quality or Just Make the Questionnaire Longer?” In *Annual Conference for the American Association for Public Opinion Research*. Austin, TX.
- Rogers, Theresa F. 1976. “Interviews by Telephone and in Person Quality of Responses and Field Performance.” *Public Opinion Quarterly* 40 (1): 51–65.
<https://doi.org/10.1086/268267>.
- Schober, Michael F., and Frederick G. Conrad. 1997. “Does Conversational Interviewing Reduce Survey Measurement Error?” *The Public Opinion Quarterly* 61 (4): 576–602.

- . 2008. “Survey Interviews and New Communication Technologies.” In *Envisioning the Survey Interview of the Future*. Hoboken, NJ: Wiley.
- Schober, Michael F., Frederick G. Conrad, and Scott S. Fricker. 2004. “Misunderstanding Standardized Language in Research Interviews.” *Applied Cognitive Psychology* 18 (2): 169–88. <https://doi.org/10.1002/acp.955>.
- Schober, Michael F., Anna L. Suessbrick, and Frederick G. Conrad. 2018. “When Do Misunderstandings Matter? Evidence From Survey Interviews About Smoking.” *Topics in Cognitive Science* 10 (2): 452–84. <https://doi.org/10.1111/tops.12330>.
- Singh, Anne-Marie, Nadine Marcus, and Paul Ayres. 2012. “The Transient Information Effect: Investigating the Impact of Segmentation on Spoken and Written Text.” *Applied Cognitive Psychology* 26 (6): 848–53. <https://doi.org/10.1002/acp.2885>.
- Suessbrick, Anna, Michael F. Schober, and Frederick G. Conrad. 2000. “Different Respondents Interpret Ordinary Questions Quite Differently.” In *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 907–12. Alexandria, VA: ASA. http://ww2.amstat.org/sections/SRMS/Proceedings/papers/2000_155.pdf.
- Sweller, John, Paul Chandler, Paul Tierney, and Martin Cooper. 1990. “Cognitive Load as a Factor in the Structuring of Technical Material.” *Journal of Experimental Psychology: General* 119 (2): 176.
- Tourangeau, Roger, Frederick G. Conrad, Zachary Arens, Scott Fricker, Sunghye Lee, and Elisha Smith. 2006. “Everyday Concepts and Classification Errors: Judgments of Disability and Residence.” *Journal of Official Statistics* 22 (3): 385.
- Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. Cambridge University Press.
- Tourangeau, Roger, and Tom W. Smith. 1996. “Asking Sensitive Questions the Impact of Data Collection Mode, Question Format, and Question Context.” *Public Opinion Quarterly* 60 (2): 275–304.
- US Census Bureau. 2019. “Educational Attainment in the United States: 2018.” The United States Census Bureau. 2019. <https://www.census.gov/data/tables/2018/demo/education-attainment/cps-detailed-tables.html>.
- White, Sheida. 2010. *Understanding Adult Functional Literacy: Connecting Text Features, Task Demands, and Respondent Skills*. New York: Routledge.
- Yan, Ting. 2005. “Gricean Effects in Self-Administered Surveys.” Unpublished doctoral dissertation, College Park, MD: University of Maryland.
- Zhang, Chan, and Frederick Conrad. 2014. “Speeding in Web Surveys: The Tendency to Answer Very Fast and Its Association with Straightlining.” In *Survey Research Methods*, 8:127–135.